ECMWF COPERNICUS REPORT

Copernicus Atmosphere Monitoring Service

Annual report on the verification of interim reanalyses

IRA2019

Issued by: METEO-FRANCE / G. Collin

Date: 31/07/2020

Ref:

CAMS50_2018SC2_D5.3.1-2019_202007_Annual_verification_report_IRA2019_v1

IMPLEMENTED BY ECMWF

This document has been produced in the context of the Copernicus Atmosphere Monitoring Service (CAMS).
The activities leading to these results have been contracted by the European Centre for Medium-Range Weather Forecasts, operator of CAMS on behalf of the European Union (Delegation Agreement signed on 11/11/2014). All information in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability. For the avoidance of all doubts, the European Commission and the European Centre for Medium-Range Weather Forecasts has no liability in respect of this document, which is merely representing the authors view.

Contributors

INERIS
F. Meleux
B. Raux
A. Ung
A. Colette

METEO-FRANCE
G. Collin
N. Assar

Table of Contents

Table of Figuresaa

Executive summary

The present report provides a performance analysis of the Regional interim air quality reanalyses throughout Europe, produced by CAMS for the year 2019.

The CAMS Regional services include the provision of ENSEMBLE air quality reanalyses, resulting from the combination of seven well-validated and documented chemistry-transport models' results. So-called "interim" reanalyses are data assimilated fields of air pollutant concentrations, based on up-to-date observation data. Since October 1st, 2015, according to EU Decision 2011/850/EU *on reciprocal exchange of information and reporting on ambient air quality*, EU Member States must report to the European Environment Agency (EEA) observation data as soon as it is produced, even if the necessary validation process is not completed. Such data is thus flagged as "non-validated" or "non-verified" data. Up-to-Date (UTD) data should be considered as provisional or "interim" data, until they are flagged as "validated" by the Member States, which can formally happen more than one year after their production[1].

Nevertheless, it is interesting to elaborate interim reanalyses as first guess of air pollution patterns and levels that developed in Europe in 2019. Such information can be used to support Member States for the regulatory reporting duty on air quality (according to Directive 2008/50/EC). This is the reason why it is important to carefully evaluate the simulations against observations that are not used for the reanalyses production.
INERIS performed this evaluation process and computed several performance indicators and scores for ozone, nitrogen dioxide, sulfur dioxide, PM_{10} and $PM_{2.5}$ concentrations. They are presented in this report relying on a list of background stations not assimilated by the models. Globally, the models performed as expected and the ENSEMBLE median reanalysis generally gives good results but not always the best ones, especially when analyzing the capability of the reanalyses to detect threshold exceedances for O_3 and PM_{10}. Consistency with previous validated interim reanalyses results is ensured.
The interim reanalyses maps can be considered as relevant for policy support, even if some care should be taken, as usual with provisional results.

We can highlight the following points:

- As for the previous years, too little up-to-date observation data was available to perform an extensive evaluation of interim reanalyses over the whole of Europe (except for Central and Western Europe). Very few observations can be available in Eastern Europe and also in the Southern and Northern European regions that are not correctly covered. This is frustrating since they correspond to areas where there are more uncertainties (especially because of emissions).
- In Western and Central Europe, where there are more stations for the evaluation of the models' performances, results are generally more representative and correct. The quality of the

[1] Validated observations related to year Y-1 are reported by September 30th of year Y by the Member States.

reanalyses is generally similar to the previous years, as slight improvement was noticed compared to 2018 about the O_3 and PM_{10} scores.

- The European Environment Agency is building capacity to strengthen quality assurance procedures in the coming years and more countries are supposed to deliver up-to-date data, which will impact positively the interim reanalyses production process.
- For all pollutants, the performances are always of lower quality than what can be achieved with the validated reanalyses process, for which more stations are available and observation datasets are validated.
- The ENSEMBLE reanalyses give the best results for ozone when focusing on classical statistical scores. Ozone daily maxima are generally underestimated. Correlation coefficient ranges between 0.8 and 0.9 and RMSE around 10 at rural and suburban locations. However, when looking at the threshold exceedances, it is worth noting the low capabilities of the ENSEMBLE reanalyses to detect concentrations above the standards (25%) and its good skills to keep the number of false alarms at a very low level. It highlights the high confidence associated with the ENSEMBLE's exceedances represented on maps.
- Excepting one model (RIUa), the model responses for ozone are very close and slightly better than in 2018 for bias, RMSE and correlation. A bigger diversity of responses appears when considering the capability of detection of the threshold exceedances.
- The performances of nitrogen dioxide ENSEMBLE reanalyses are quite stable with previous years and with satisfactory scores, but lagging behind the performances of the other pollutants (without considering SO_2). RMSE is around 12 $\mu g/m^3$, bias shows an underestimation of 10 $\mu g/m^3$ and correlation is close to 0.7. One model behaves as outlier (SMHa).
- PM_{10} is the pollutant for which model responses range in a large interval. Two models (RIUa and KNMa) are aside of a group (including the ENSEMBLE), where correlation coefficient ranges from 0.8 to 0.9 and RMSE from 4 to 6 $\mu g/m^3$, depending on the model and the station typology. Model responses have a bias close to 0 with a tendency to become slightly negative for urban stations. Anyway, the ENSEMBLE shows good performances, better than for the previous year. Part of this result might be explained by the scores over Polish stations which improved this year. The homogeneity of the ENSEMBLE's scores whatever the typology of stations considered is also noticed.
- Moreover, the evaluation demonstrates how the Ensemble approach, based on a median average of involved models, is not appropriate to simulate exceedances of threshold values. Only 35% of good detection of exceedances of the PM_{10} daily limit values was correctly caught by the ENSEMBLE, whereas the best reanalyses got 60%. As for ozone, the ENSEMBLE reanalyses produce a very low number of false alarms.
- Although only little $PM_{2.5}$ measurement data was available for the evaluation, the results obtained for this pollutant are promising. The individual models' responses are quite consistent, and the Ensemble median gives the best results. Correlation coefficient is close to 0.9 and RMSE between 4 and 5 $\mu g/m^3$, which is good. Once again, the conclusions are limited by the low number of stations available in some geographical areas and should be consolidated and improved in future interim assessments, when the up-to-date data gathering process at the EEA is strengthened.
- The model representativeness is limited to correctly reproduce SO_2 concentrations in Europe, due to the characteristics of the emissions sources of such pollutant. This is illustrated by the

results that show low RMSE and bias due the low background concentrations measured and very poor correlation for almost all European stations, highlighting the complexity for the model to reproduce the temporal variability of the concentrations.

Introduction

This report gives an overview of the performances of the European air quality **interim reanalysis** process developed by the CAMS Regional services and implemented to simulate air quality in Europe during the year 2019.

Air quality interim reanalyses result from a combination of chemistry-transport models' results that simulate the spatio-temporal evolution of regulatory air pollutant concentrations (according to the ambient Air quality Directive 2008/50/EC), and observations assimilated in each model to correct and improve its results. Each team providing air quality reanalyses developed appropriate and validated data assimilation chains to provide best estimates of air pollution patterns according to available observation data.
The models implemented to calculate these interim reanalyses are the set of seven models run in other near-real-time CAMS Regional services. The models are CHIMERE (INERIS, France), EMEP (MET Norway, Norway), EURAD-IM (FZJ-IEK8, Germany), LOTOS-EUROS (KNMI-TNO, The Netherlands), MATCH (SMHI, Sweden), MOCAGE (METEO-FRANCE, France), and SILAM (FMI, Finland).
Observations are issued from the regulatory air quality monitoring networks that report to the European Environment Agency (EEA), according to Air Quality Directive 2008/50/EC and Decision 2011/850/EU on reciprocal exchange and reporting on ambient air quality. "Interim reanalyses" are so called because the observation data used are not formally validated yet. The 2011 decision stipulates that Member States must report monitoring data as soon as they are produced, in near-real-time, with an appropriate flag indicating that they are not verified or validated yet. This set of data is named "Up-To-Date (UTD) data". The data is gathered in the commonly named AQ e-reporting database. "Interim data" are UTD data collected on the EEA website within a certain delay, to leave enough time to have a chance to get verified data[2]. We estimate that 20 days is appropriate time-lag to get the data and run the reanalyses for a given day.

The set of observation sites reported to the EEA is split into two subsets, one for data assimilation with almost 2/3 of the stations) in the interim reanalyses and the other (the remaining 1/3 of stations) for verification. Those datasets do not overlap, and verification cannot be biased by use of data for both assimilation and verification processes. It should be noted that not all Members States reported UTD data and other countries just start (like Italy) with partial observed dataset made available. Consequently, data assimilation and evaluation cannot be performed in some geographical areas and the robustness of the results may vary from one area to another. Therefore, it will not be possible to draw some clear conclusions about the model capacities in those regions.

The evaluation focuses on the seven individual models and the ENSEMBLE as well. The ENSEMBLE is the result of the median of the seven models and is considered as the best estimate of air pollution

[2] Member states can check, verify and validate their data when they want and resubmit with the appropriate flag as many times as they wish. Formal validation is expected only in September the year after.

patterns and levels, since it combines the strengths of the other models. This is what will be checked in the present report.
Statistical indicators (bias, root mean square error, correlation coefficient) are presented to compare the models' results against observations. Maps, histograms and Taylor diagrams are proposed for a better understanding and analysis of the performances. They are computed for the four regulatory pollutants targeted by the service: ozone (O_3), nitrogen dioxide (NO_2), particulate matter (PM_{10} and $PM_{2.5}$). Metrics relevant for policy purposes (regarding the content of the air quality directives) and for health impacts are considered for the evaluation.
All results are presented below, after a short introduction on the computed performance indicators.

1. Performance indicators

The model performances are evaluated on the basis of classical statistical indicators which measure objectively the gap between the model results and the observations at the available stations: bias, root mean square error (RMSE) and correlation coefficient are the most classical. Comparison of observed and modelled averages is generally considered as well.
Obviously, the behavior of performance indicators depends on the station typology and the considered pollutant: the models used in the CAMS Regional service run at the European scale and their spatial resolution is about 10 km. Consequently, for pollutants which are largely influenced by local sources (NO_2, PM in some situations), these regional models are not able to reproduce hot spots monitored by traffic or industrial stations, and performance indicators will not be assessed. Difficulties can even be encountered at urban stations.

Conversely for pollutants characterized by long residence time in the atmosphere and large impacted areas (typically ozone and PM in some cases), performance indicators evaluated at all type of stations (except traffic and industrial sites) make sense.

The definitions of the various performance indicators used in the report are given below. They are very usual[3] in evaluation processes:

- Bias indicates, on average, if the simulations under or over-predict the actual measured concentrations. In our case, negative values indicate under-prediction, whereas positive values indicate over-prediction; values close to 0 are the best ones:

  $$\frac{1}{N}\cdot\sum_{i=1}^{N}(P_i - O_i)$$

  Where N is the number of observations, P_i refers to the predictions and O_i to the observations. It is expressed in µg/m^3.
- Root Mean Square Error (RMSE) gives information about the skill of the model in predicting the overall magnitude of the observations. It should be as weak as possible:

  $$\sqrt{\frac{1}{N}\cdot\sum_{i=1}^{N}(P_i - O_i)^2}$$

  Where N is the number of observations, P_i refers to the predictions and O_i to the observations. It is expressed in µg/m^3.

- Correlation is a measure of whether predictions and observations change together in the same way (i.e. at the same time and/or place). The closer the correlation is to one, the better is the correspondence of extreme values of the two data sets.

  $$r = \frac{\mathrm{cov}(P_i, O_i)}{\sqrt{\mathrm{var}(P_i)}\cdot\sqrt{\mathrm{var}(O_i)}}$$

[3] Chang J.C. et Hanna S.R., 2004. Air quality model performance evaluation. *Meteorol. Atmos. Phys*. 87, 167–196.

Where N is the number of observations, P_i refers to the predictions and O_i to the observations. This is a non-dimensional number.

Taylor diagrams synthesize on a unique quadrant, various statistical indicators for different models: the radii correspond to the correlation coefficient values, the x-axis and the y-axis delimits arcs with bias values and the internal semi-circles correspond to the RMSE values. Therefore, this is a very pedagogic way to present an overview of the relative performances of a set of models, often used in model intercomparison exercises.

For indicators related to threshold values, for instance the number of days, hours when a certain concentration level is exceeded, some 'contingency tables' giving the percentages of correct predictions (GP), false alarms (FA), or missing events (ME) are estimated. These concepts come from the weather or air quality forecasting world. Although they are very severe and not objectively representative of the intrinsic model performance (because of the threshold cut-off effect, a result close to the threshold can fall arbitrary in one or the other category), they can give useful information to compare various models' behaviors in different geographical regions. GP, FA and ME are expressed in percentage (%) and also referred sometimes to the total number of stations within each class (GP, FA and ME). Based on these values, several ratios are defined providing various information about the model:

- Probability of good detection POD= GP / (GP+ME)
- Success ratio SR = GP / (GP+FA)
- Critical success index CSI = GP / (GP +ME +FA)
- Fbias = (GP + FA) / (GP + MA) – which represents whether the model tends to overestimate (FBIAS > 1) or underestimate (FBIAS < 1) the threshold exceedances.

Several representations of the models' skills are proposed:

- Maps with colored patches at the location of the stations selected for the evaluation process. The color scale indicates how the model performs.
- Taylor diagrams provide a wider overview of the model performances.
- Histograms with model performances sorted by station typology and by European sub-region (Western, Northern, Southern, Central, Eastern) are proposed as well.
- Performance diagram giving an overview of the skills to detect threshold exceedances, by plotting together on the same figure POD, SR, CSI and Fbias. The objective for the models is to be the closest of the upper right corner.

2. Performance indicators for ozone

In this evaluation, we focused on the ability of the model to correctly represent the ozone daily maximum (hourly average), which is the most relevant considering regulatory indicators like the number of exceedances of information and alert thresholds. The evaluation is performed over the "summer" period when ozone increases, reaching levels that may impact human health and ecosystems.

Figure 1 shows the Taylor diagram that synthesizes performances of individual CAMS models and the ENSEMBLE to simulate hourly daily maximum of ozone in the summer period. The scores in this figure are computed with all typologies of background stations. The graph shows similar performances for most of the reanalyses (one is aside of the main group), highlighting slightly better scores for the ENSEMBLE with correlation slightly above 0.9 and RMSE around 9 µg/m^3.

In-depth analysis of the interim ENSEMBLE reanalyses can be elaborated considering the spatial distribution of the statistical indicators over Europe. Figure 2 presents maps of bias, correlation coefficient and RMSE related to the ENSEMBLE, for daily maxima from the 1st April to 30th September 2019. Bias ranges in most parts of Europe between -5 and 5 µg/m^3. Most of the stations show an underestimation of the ozone concentrations, which led to an overall bias of -4 µg/m^3. However, it should be noted that evaluation cannot be conducted in several Southern countries (Greece, Serbia) and Eastern countries (Romania, Bulgaria), because of a lack of reported interim observation data.
Correlation coefficient is excellent with high values, most of them higher than 0.9. The same quality can be seen with RMSE, most of the scores are below 15 µg/m^3, although higher values are found along the Mediterranean area (Italy, Slovenia, Croatia and Spain) up to 25 µg/m^3.
This can be a consequence of using partial and non-validated observation data, and results should improve when the validated reanalyses are performed. However, results remain acceptable compared to the state of the art. It is worth noting better performances compared to one year before, with an improvement of the RMSE from 12 to 9 µg/m^3, a slightly better correlation and stable bias. It seems that part of the improvement occurred over stations from Central Europe (Slovakia, Hungary and Poland).

To help in the interpretation of those maps, one can consider the same performance indicators for each individual model and the ENSEMBLE and various station typologies. Figure 3 and Figure 4 present bias, RMSE, and correlation coefficient scores for all models, at rural and suburban stations respectively. The indicators are sorted per geographical region: Western Europe (EUW), Central Europe (EUC), Southern Europe (EUS), Northern Europe (EUN), Eastern Europe (EUE). The interpretation of the results is hampered by the low number of stations available for verification in some areas (in Northern and Eastern Europe). The number of stations taken in consideration for computing the scores is mentioned on the figures. In Eastern Europe, no station was available for suburban site scores; the verification process has not been performed since very few countries in that area report UTD observation data to the European Environment Agency. The situation is expected to improve in the coming years, like it has improved in this report for Central Europe and Southern countries compared to previous years with the integration of more UTD observations

from Italy. In Eastern and Northern Europe, the evaluation has been performed against a very low number of stations, which may be a problem regarding the representativeness of the obtained results.

Where observation data is available, the panel of reanalyses shows common underestimations in the scores over Western Europe, Central Europe and Southern Europe. Model responses are similar between these three areas, with a more pronounced underestimation over Southern countries.
The underestimations show more variability when focusing on suburban stations in Western and Central Europe, while overestimations appear in Southern suburban stations.
Regarding RMSE, we can note once again good consistency between results, with an increase of the values in Southern countries compared to Western and Central rural stations. The scores over suburban stations in Central Europe are better than in Western Europe.
Correlation coefficient is quite high ranging from 0.8 to 0.9 in best cases, among which is the ENSEMBLE. Only one reanalysis has scores aside of the group with correlations close to 0.5 and far from the best ones.
Obviously, there are more uncertainties in the models in Southern, Eastern and Northern regions, due to uncertainties in emissions and the complexity of the photochemical processes and meteorology. However, there are also much fewer stations than in other regions, making the scores very sensitive to the weak performance of one or two stations. For this reason, conclusions should be established with care and refined when validated reanalyses for 2019 are available.
Nevertheless, overall performances of the models to simulate ozone daily maxima are satisfactory and consistent with previous results obtained in the past and with the state of the art.

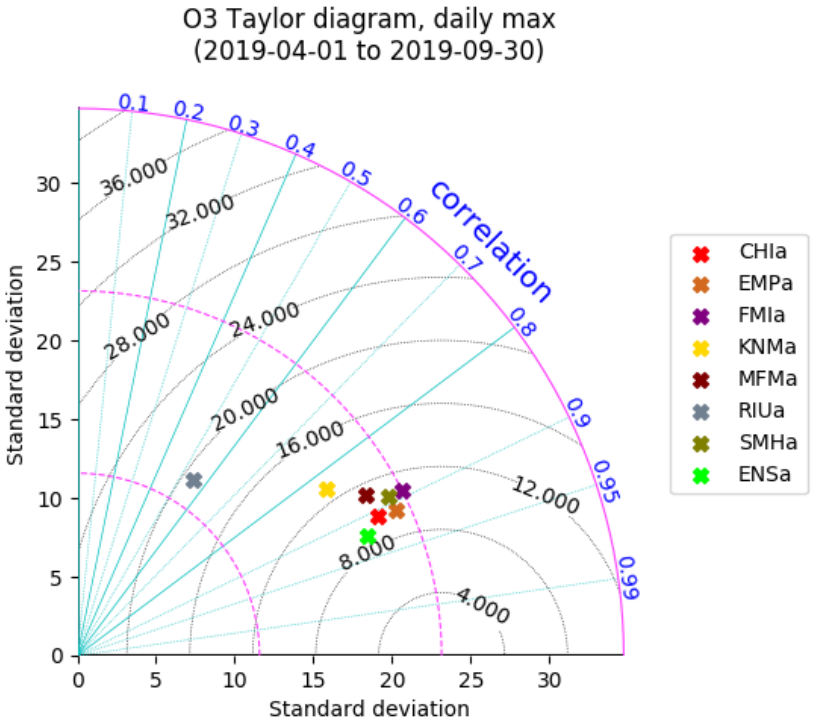
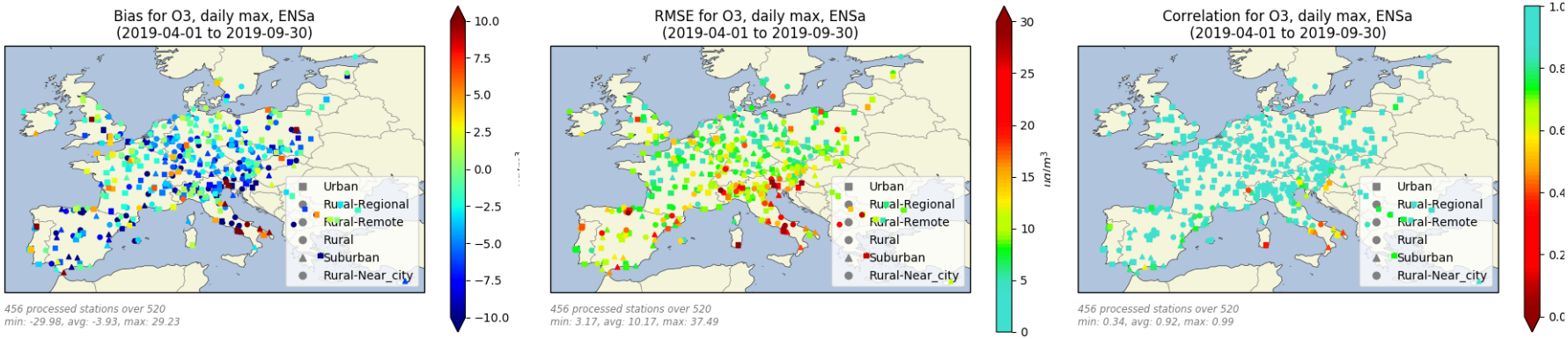
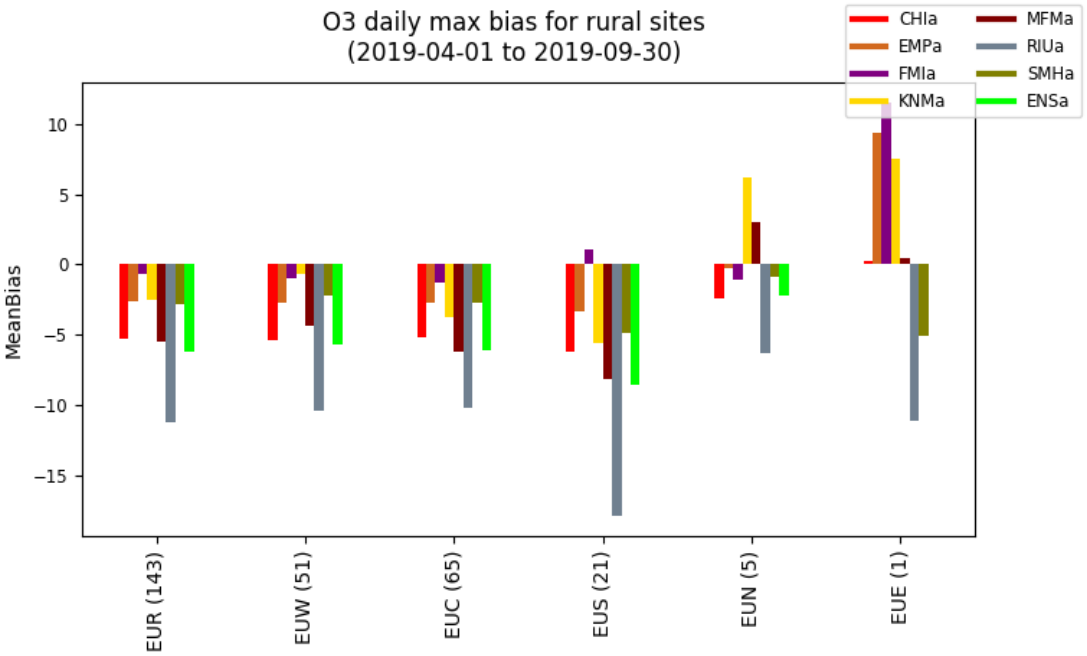


Figure 1 - Taylor diagram presenting performances of all CAMS regional models to simulate summer ozone daily maximum (hourly average).

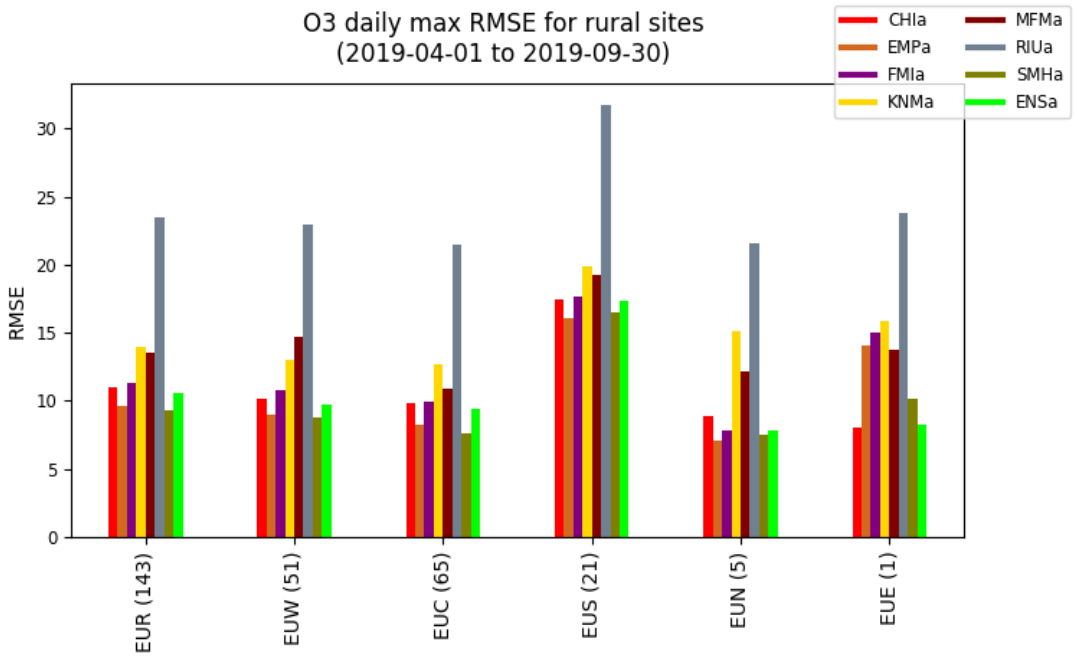


(a) ozone bias (daily maximum) (b) ozone RMSE (daily maximum) (c) ozone correlation coefficient (daily maximum)

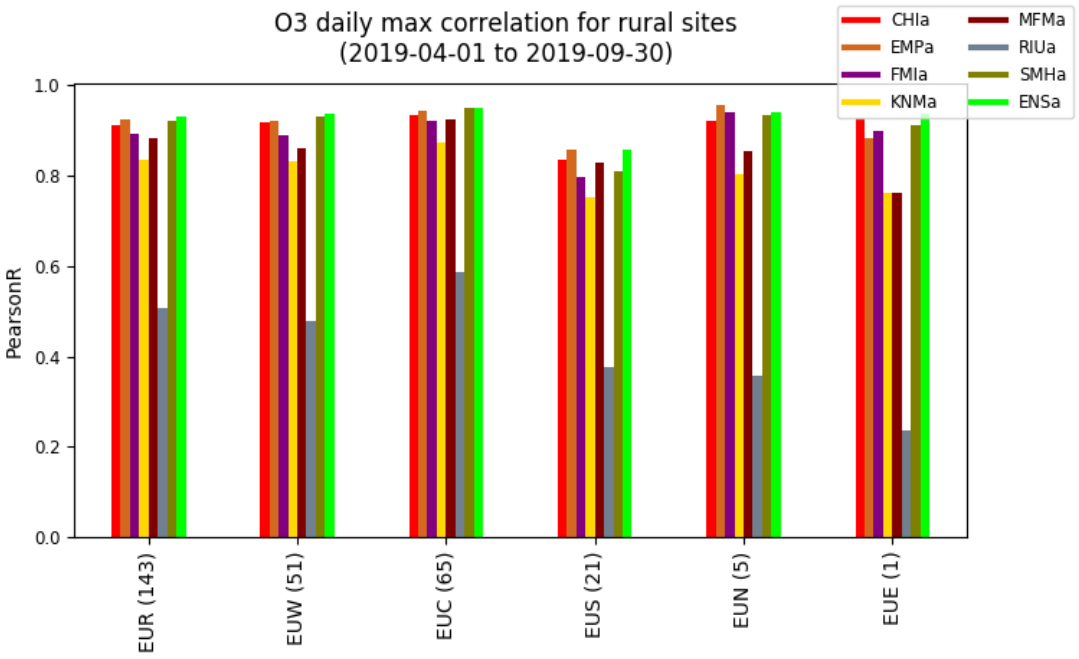
Figure 2 - Maps of Statistical scores of the ENSEMBLE interim reanalyses results against the observation validation dataset from the AQ e-reporting database for the ozone daily maximum, from 01/04/2019 to 30/09/2019: (a) Bias, (b) RMSE, (c) Correlation coefficient.


O3 daily max bias for rural sites
(2019-04-01 to 2019-09-30)
CHIa
EMPa
FMIa
KNMa
MFMa
RIUa
SMHa
ENSa
MeanBias
10
5
0
−5
−10
−15
EUR (143)
EUW (51)
EUC (65)
EUS (21)
EUN (5)
EUE (1)

(a)

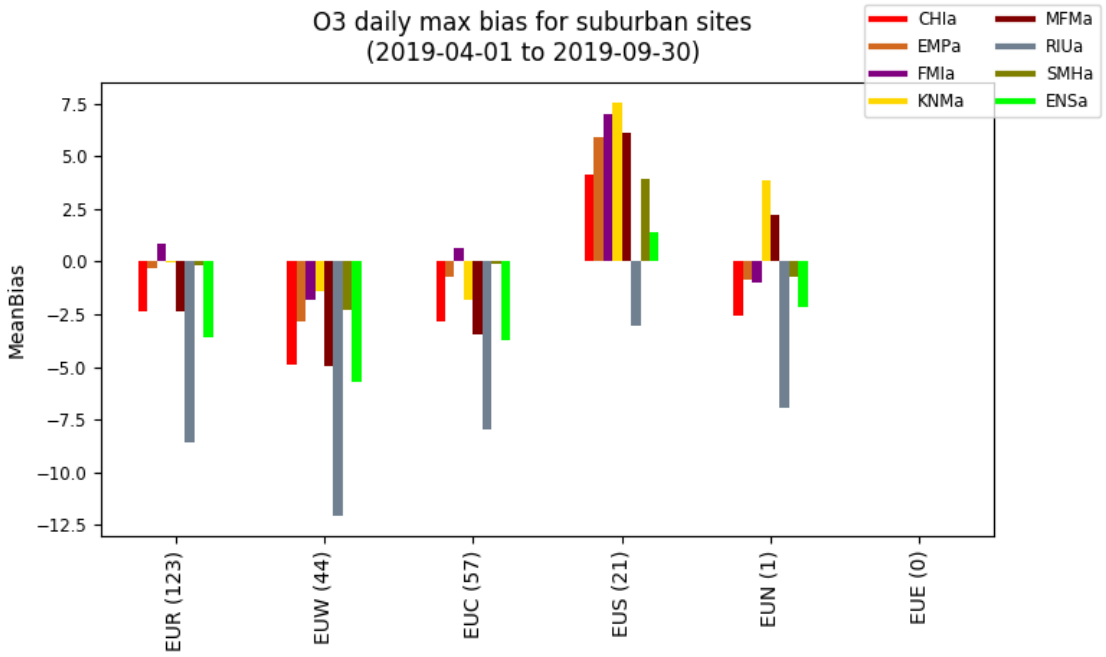

O3 daily max RMSE for rural sites
(2019-04-01 to 2019-09-30)
CHIa
EMPa
FMIa
KNMa
MFMa
RIUa
SMHa
ENSa
RMSE
30
25
20
15
10
5
0
EUR (143)
EUW (51)
EUC (65)
EUS (21)
EUN (5)
EUE (1)

(b)

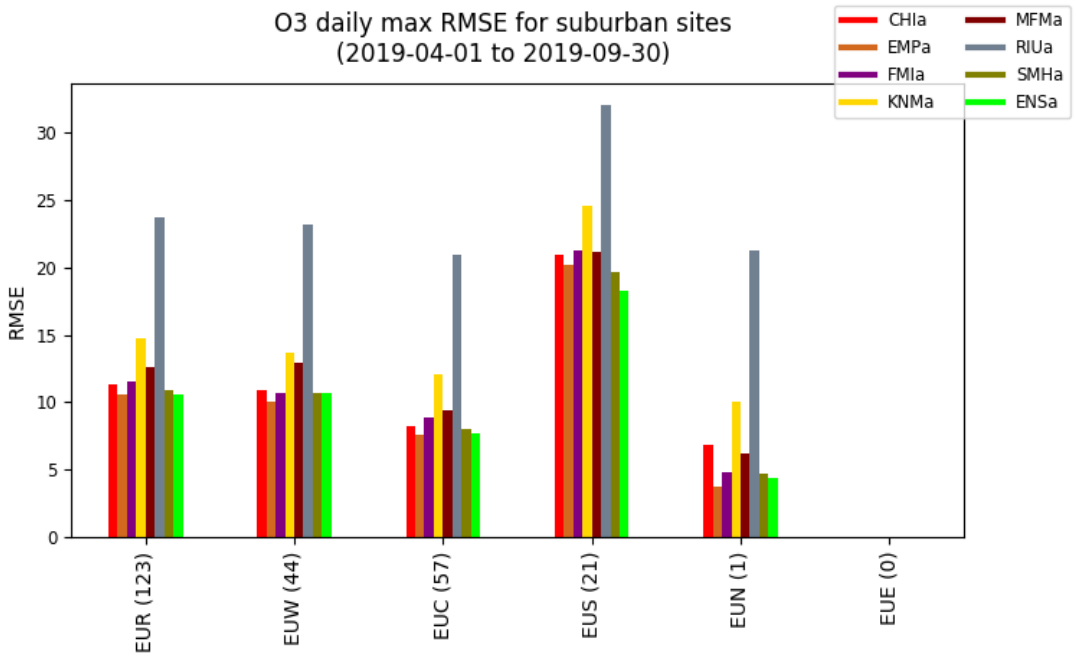


(c)

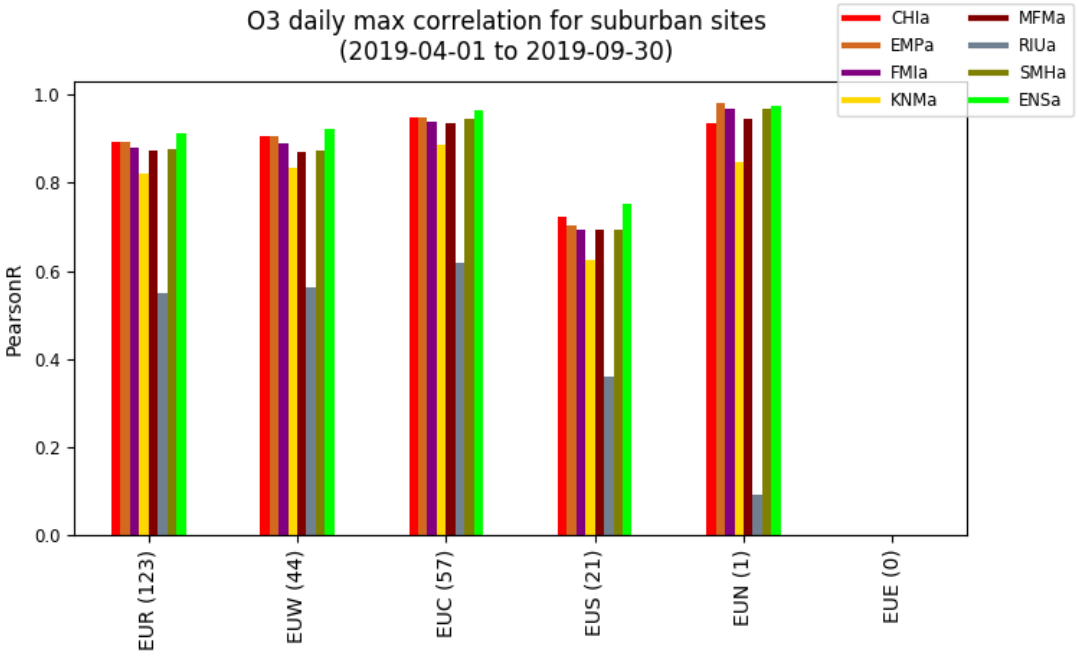
Figure 3 - CAMS Regional interim reanalyses for predicting daily ozone peak over the summer 2019 throughout European sub-regions: (a) Bias (b) RMSE (c) Correlation coefficient at rural stations.



(a)



(b)



(c)

Figure 4 - CAMS Regional interim reanalyses for predicting daily ozone peak over summer 2019 throughout European sub-regions: (a) Bias (b) RMSE (c) Correlation coefficient at suburban stations.

Finally, the models' ability to simulate the number of exceedances of a given threshold value has also been assessed. This is important for ozone, since the EU legislation (Directive 2008/50/EC) sets quality objectives with an information threshold (180 µg/m^3) and an alert threshold (240 µg/m^3), over which short-term action plans and communication towards the general public should be implemented by Member States. However, this kind of evaluation against threshold value is very stringent and not always representative of the model quality. Situations above and below the threshold value are counted, but to correctly take into account model uncertainty, it would be necessary to take a range of acceptable values around the threshold. This is not done in the present study. Therefore, the diagnosis can be seen as a pessimistic analysis of the models' performances.

Figure 5 below shows the number of situations when exceedances of the hourly information threshold have been observed during the summer time in 2019 (time is presented on the x-axis), sorted per geographical region (various colors). The first set of histograms shows observed exceedances at ozone stations in Europe. Most of the exceedances were located in Central, Western and Southern Europe. In total, 642 exceedances were recorded on the stations for verification, less than in 2018. The variability between geographical areas should be interpreted with caution, as it is highly dependent of the number of stations available.
The first episode occurred at the end of June 2019 and lasted 5 days. A second one has been recorded during the second half of July. The figure in the middle panel displays exceedances modelled by all CAMS models and the ENSEMBLE. If the performances of the ENSEMBLE were good considering statistical indicators, they show very disappointing results for threshold indicators. Less than 20 % of the exceedances were detected by the ENSEMBLE (Figure 6). This can be explained by the nature of the indicator (no range of uncertainty is taken into account), but also by the way the ENSEMBLE is built up. It is based on the median of individual model results, with performance varying largely from a model to another. The median smooths the indicator (evaluation against threshold values) and the obtained results cannot be considered representative of the actual quality and accuracy of the models. Overall performances of the ENSEMBLE also show negative bias, which usually does not help to detect the exceedances.
However, the contingency plot highlights the low number of false alarms made by the ENSEMBLE while other models which detect more exceedances have many more false alarms, therefore also highlighting a positive aspect of the conservative choice of the median.
The performance diagram shows that the ENSEMBLE does not have the highest probability of detection with 25% of good detection but the highest success ratio (90%), meaning that a good confidence can be associated with the detections computed by the ENSEMBLE. It has a large tendency to underestimate the threshold exceedances. The reanalysis with the best POD (around 50%) has a success ratio slightly above 50%.

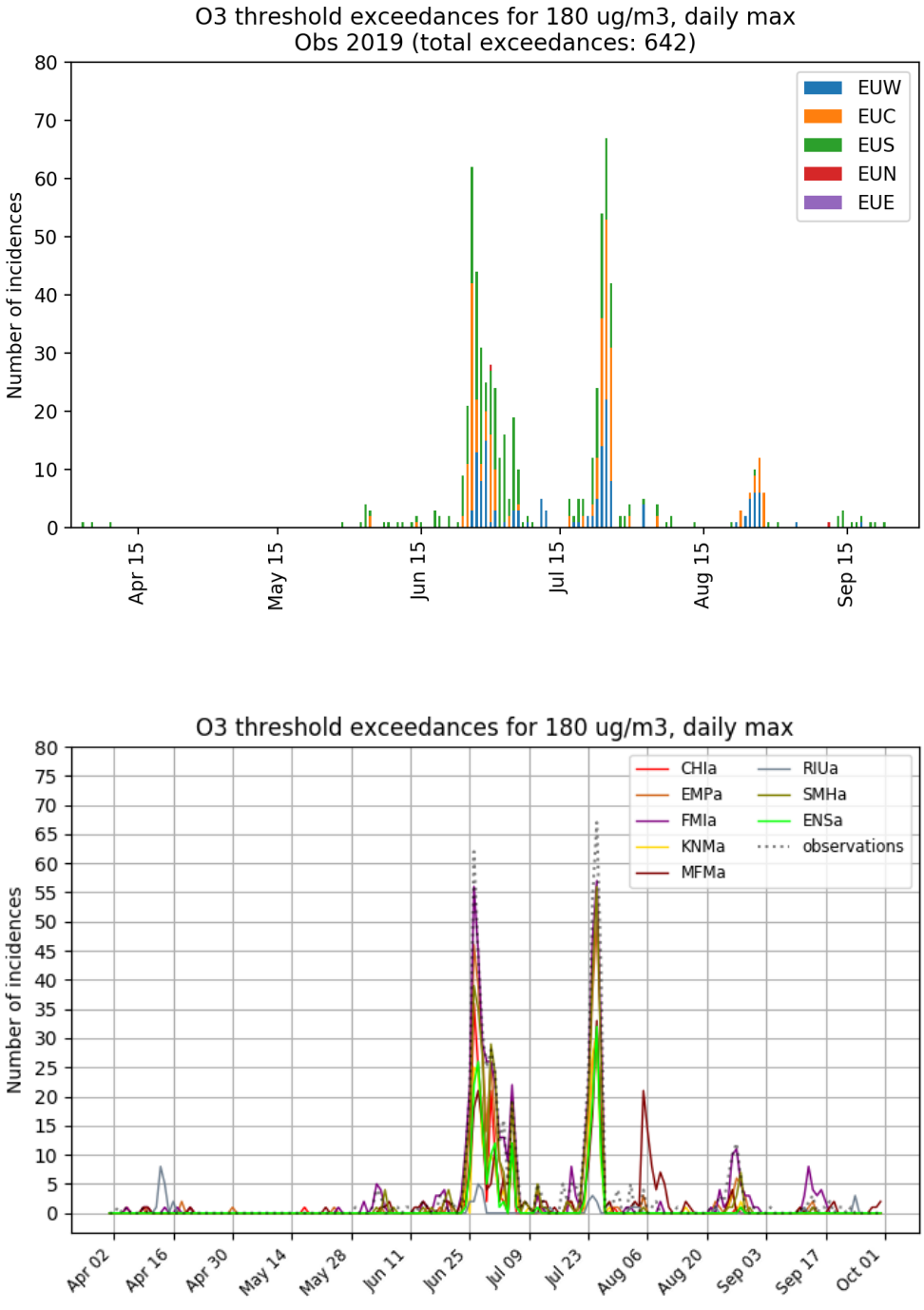


Figure 5 - Number of exceedances of the information threshold value for ozone in summer 2019 – observed (top), modelled by all the interim analyses in colour lines and observed in black dashed line (bottom).

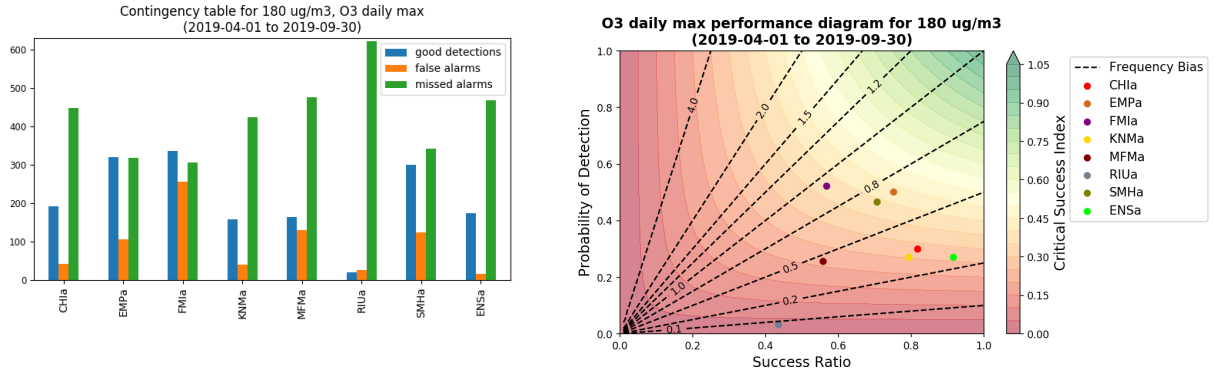


Figure 6 - Histograms describing the models performances regarding the number of exceedances of the ozone thresholds (left) and performance diagram (right).

3. Performance indicators for nitrogen dioxide

Warning note: *It should be reminded that the CAMS Regional mapping system is not fitted to deal with local hot spot situations, such as those that develop near busy roads or on industrial sites. Actually, the model resolution of 10 km is not sufficient to catch actual NO_2 concentrations at traffic and industrial sites.*

Figure 7 presents the Taylor diagram for the CAMS Regional interim reanalyses of the ENSEMBLE and its members, for the daily maximum (hourly average) of NO_2 concentrations. It shows disperse model performances; among them the ENSEMBLE has one of the best ones, with correlation close to 0.7 and RMSE around 11 µg/m^3. The worst performances depicted on this diagram are a correlation of 0.5 and RMSE around 14 µg/m^3. It is worth noting that such scores are slightly better than the scores obtained in 2018 (Figure 8), with less disperse individual performances. Maps in Figure 9 allow highlighting a tendency to underestimate NO_2 daily maximum throughout Europe, even if some isolated stations show overestimations. The RMSE is in 2019 similar to those of 2018. Same conclusions are drawn for the correlation: an overall correlation of 0.7, with highest values for some stations located in Italy, Spain and Poland.

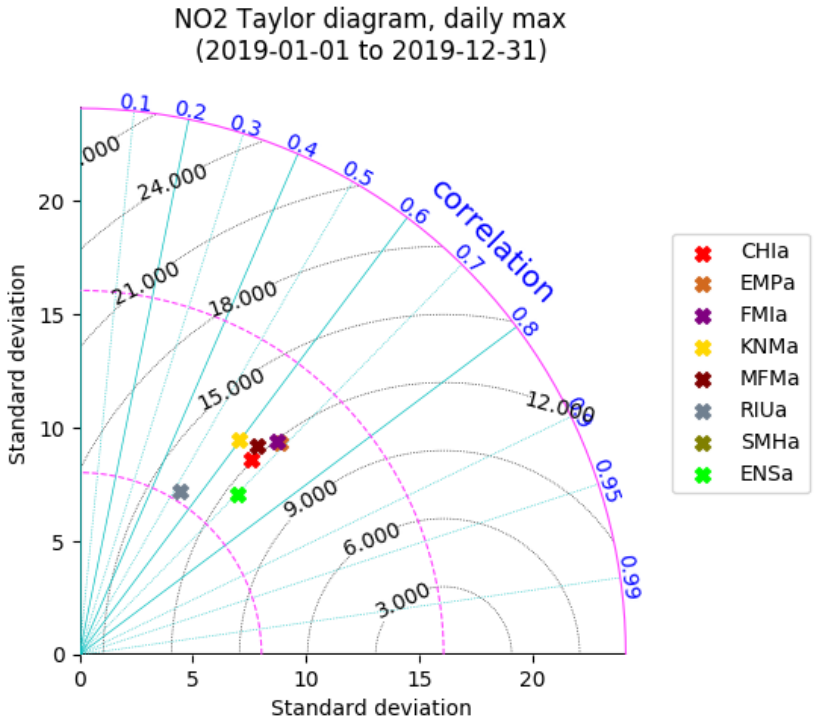


Figure 7 - Taylor diagram presenting the performances of the CAMS Regional interim reanalyses to predict NO_2 daily maxima in 2019.

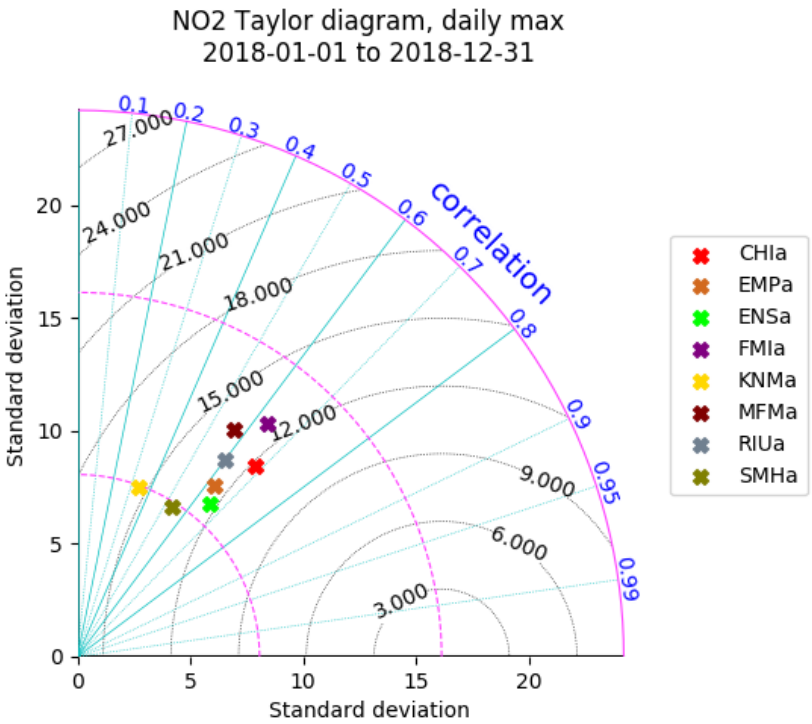


Figure 8 - Taylor diagram presenting the performances of the CAMS Regional interim reanalyses to predict NO_2 daily maxima in 2018.

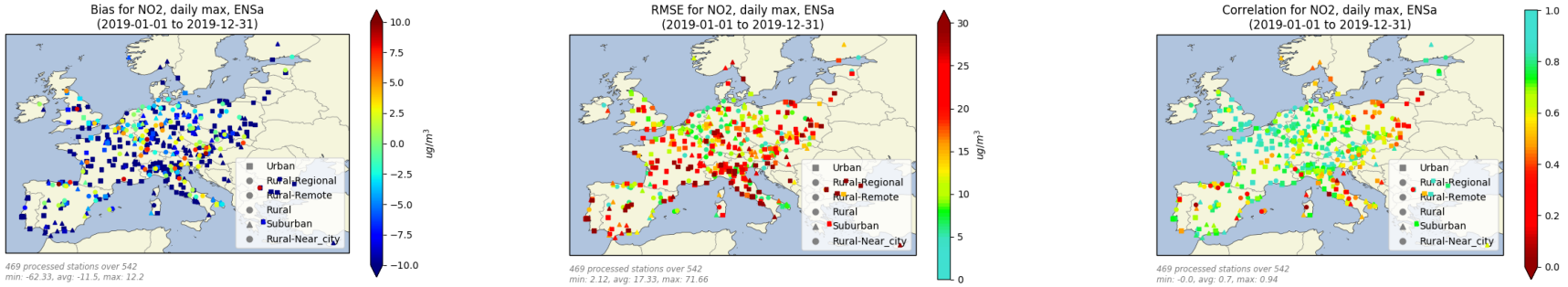


Figure 9 - Maps of Statistical scores of the ENSEMBLE interim reanalyses results against the observation validation dataset from the AQ e-reporting database for the NO_2 daily maximum over the year 2018: Bias (a) RMSE (b), Correlation coefficient (c).

4. Performance indicators for PM_{10}

Figure 10 shows the Taylor diagram obtained for PM_{10} daily averages over the year 2019, for CAMS regional individual and ENSEMBLE reanalyses. The results are very encouraging with for the ENSEMBLE a correlation coefficient between 0.9 and 0.95, which is surrounded by 5 other models with also high performances compared to the previous years. Two reanalyses performances lie out of the main group, with RMSE around 8 µg/m^3 and correlation of 0.5 for one and 0.65 for the other. ENSEMBLE RMSE is around 4 µg/m^3, which is better than what was obtained for the 2018 reanalyses (6 µg/m^3).

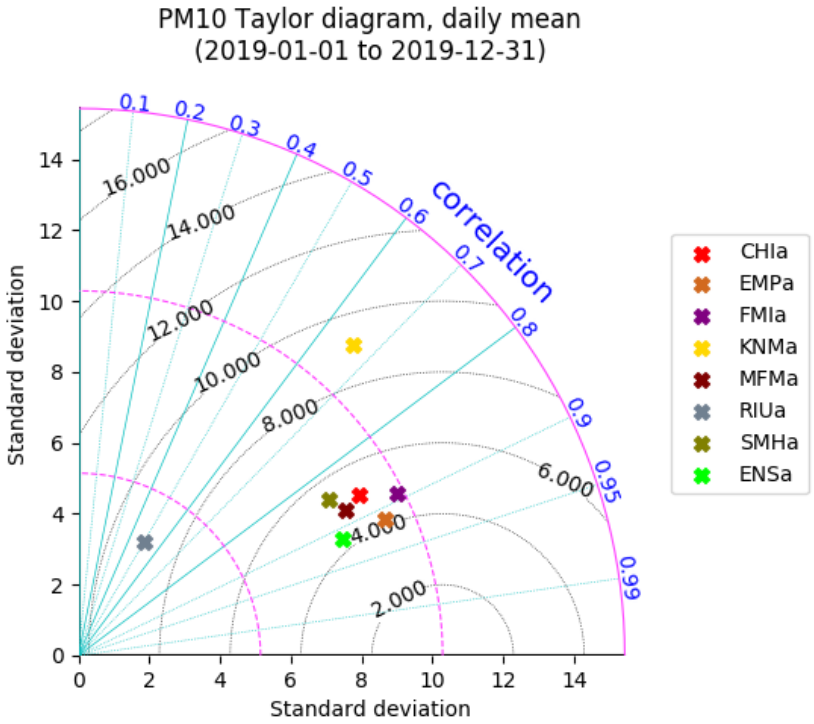
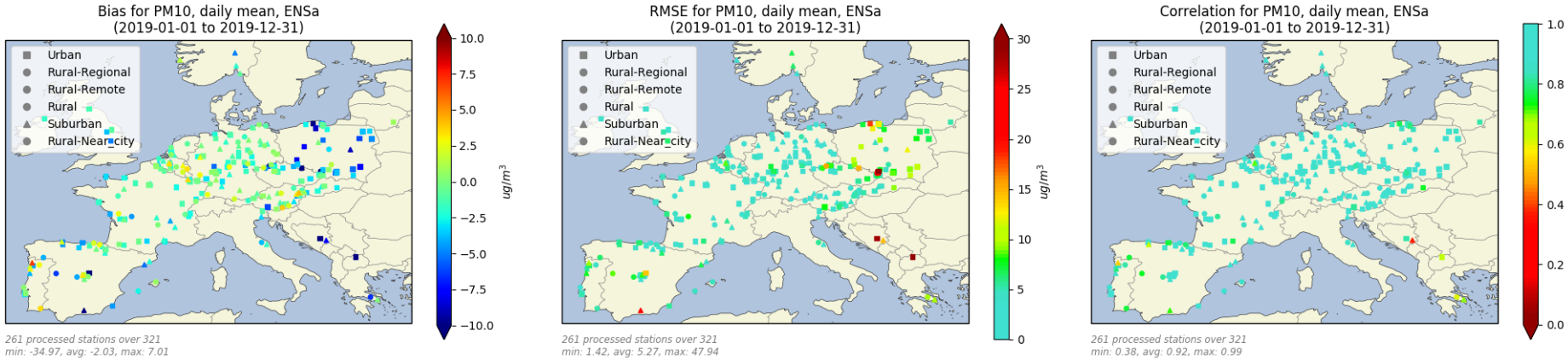


Figure 10 - Taylor diagram presenting the performances of the CAMS Regional ENSEMBLE interim reanalyses to predict PM_{10} daily average in 2019.

Figure 11 details the geographical distribution of statistical scores (bias, correlation coefficient and RMSE), for the ENSEMBLE interim reanalyses for the year 2019. Lowest scores (bias, correlation and RMSE) are obtained for stations located in Poland and in the Central-Eastern parts of Europe. In several countries (France, Germany, Benelux, the UK), RMSE ranges between 1 and 5 µg/m^3, which remains very good and similar to past year performances. The overall score of the ENSEMBLE is -2 µg/m^3, even if an almost null bias is frequent over European stations for the ENSEMBLE. The RMSE looks more homogeneous this year than in 2018, thanks to the Polish stations that have values closer to those of the other countries. Still, some isolated stations in Serbia and Macedonia show poor performances.

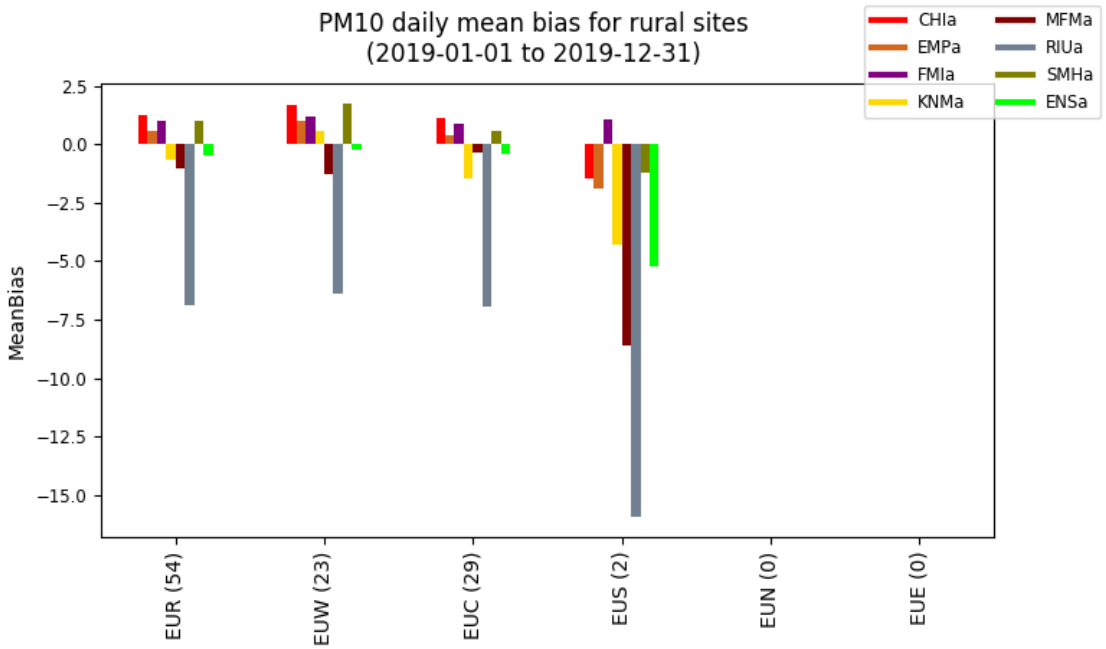
Furthermore, differences between model results can be investigated considering histograms of scores per region and for each model. Figure 12, Figure 13 and Figure 14 show these results for rural, suburban and urban stations respectively. They confirm the low number of stations available for the verification of interim PM_{10} reanalyses, with huge gaps in some areas (Southern, Northern and Eastern Europe for all station typologies). Once again, results are only robust over Western and Central Europe. For these two areas, ENSEMBLE performances are quite similar with a bias close to 0 over rural and suburban stations and slightly underestimating them over urban stations. The model responses show little variability around 0, except one model aside of the group which underestimates the PM_{10} concentrations significantly.
RMSE gives similar values whatever the typology considered. ENSEMBLE performance is among the best. Model responses including ENSEMBLE are very close for 6 out of 8 reanalyses, with values around 5 $\mu g/m^3$. The two reanalyses (RIUa and KNMa) apart from the group have RMSE between 10 and 15 $\mu g/m^3$.
These two reanalyses also have correlations well below the rest of the reanalyses, which reach values above 0.8 for rural stations and higher values for suburban and urban stations.

Despite the variabilities of the model responses, especially with two reanalyses well out of the group in terms of performances even if their scores remain acceptable, the ENSEMBLE is most of the time the reanalysis providing the best description of the PM_{10} distribution over Europe. The performances look slightly better than last year.

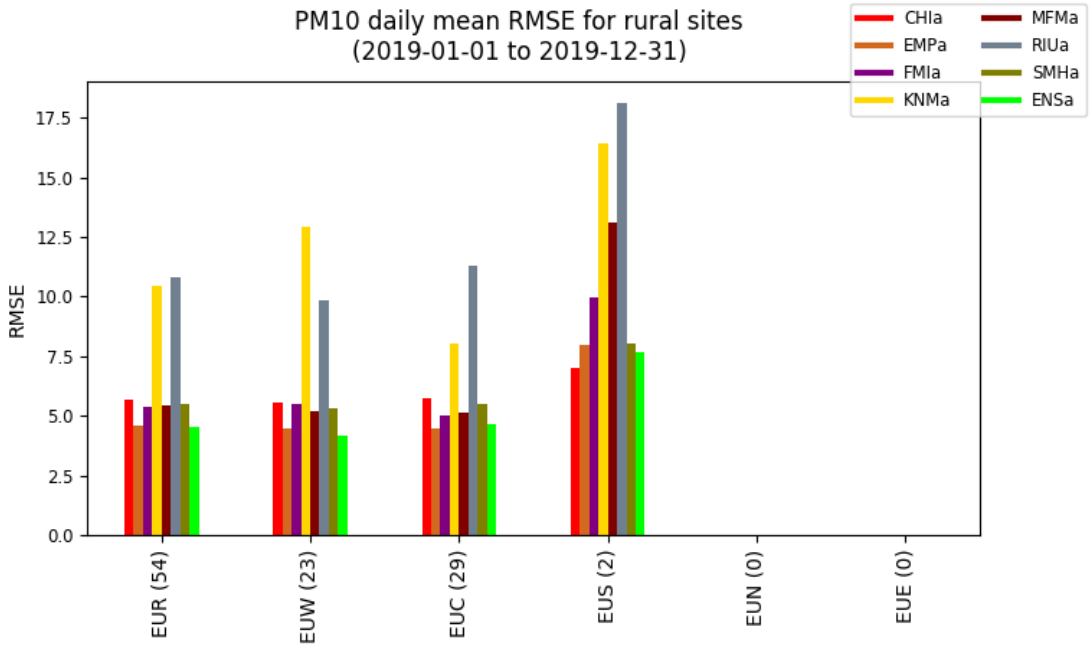


(a) (b) (c)

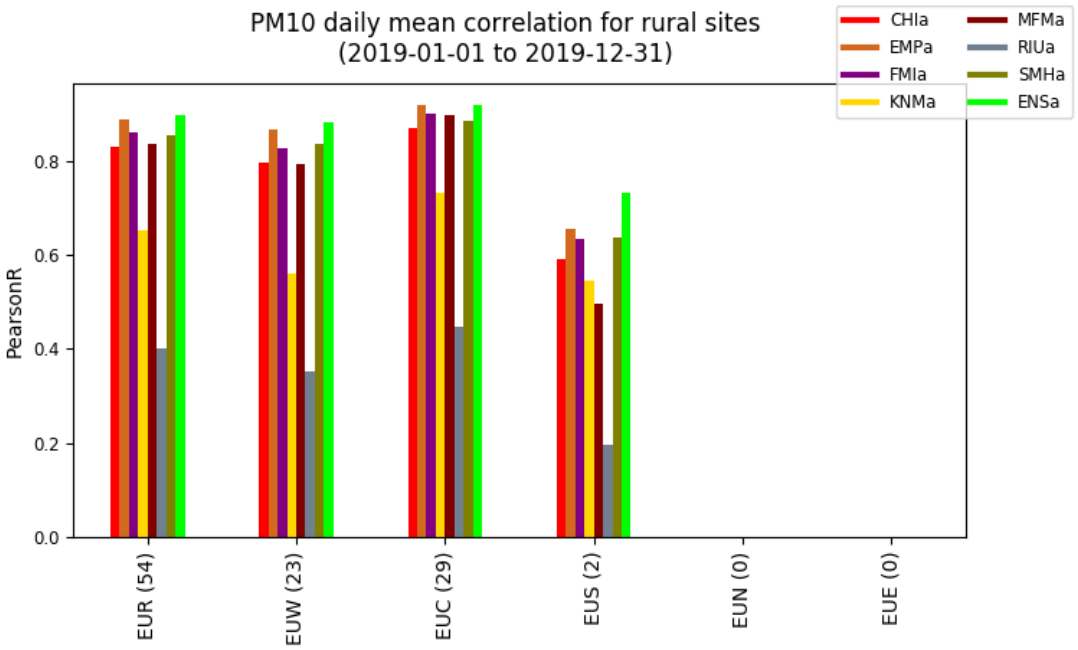
Figure 11 - Maps of Statistical scores of the ENSEMBLE interim reanalyses results against the observation validation dataset from the AQ e-reporting database for the PM_{10} daily average over the year 2019: Bias (a) RMSE (b), Correlation coefficient (c).


PM10 daily mean bias for rural sites
(2019-01-01 to 2019-12-31)
CHIa
EMPa
FMIa
KNMa
MFMa
RIUa
SMHa
ENSa
MeanBias
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−12.5
−15.0
EUR (54)
EUW (23)
EUC (29)
EUS (2)
EUN (0)
EUE (0)

(a)

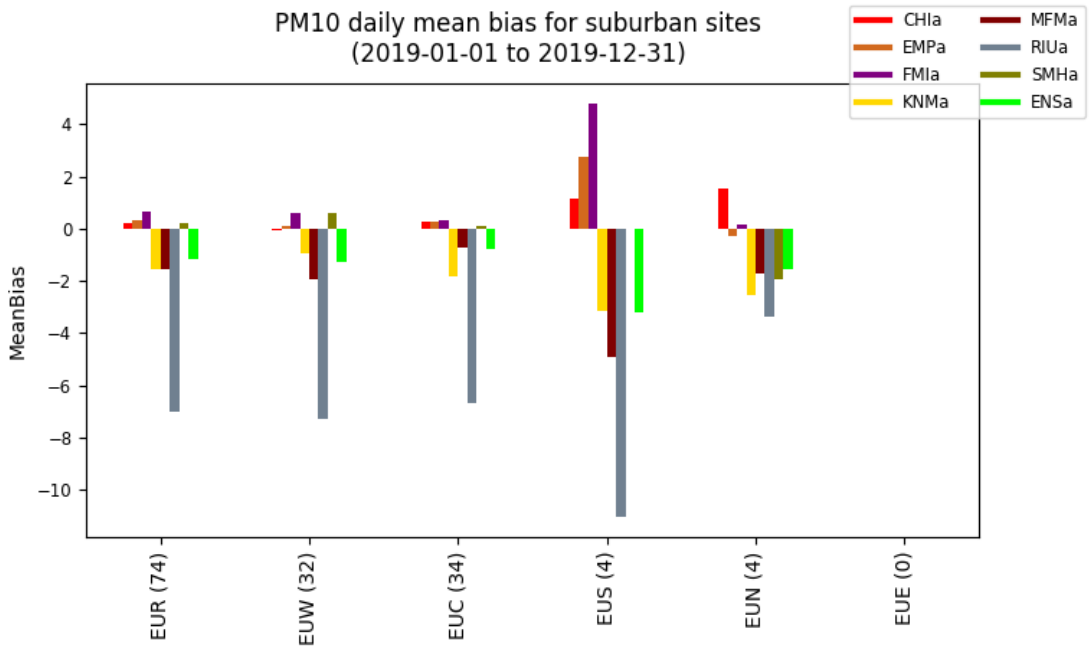

PM10 daily mean RMSE for rural sites
(2019-01-01 to 2019-12-31)
CHIa
EMPa
FMIa
KNMa
MFMa
RIUa
SMHa
ENSa
RMSE
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
EUR (54)
EUW (23)
EUC (29)
EUS (2)
EUN (0)
EUE (0)

(b)

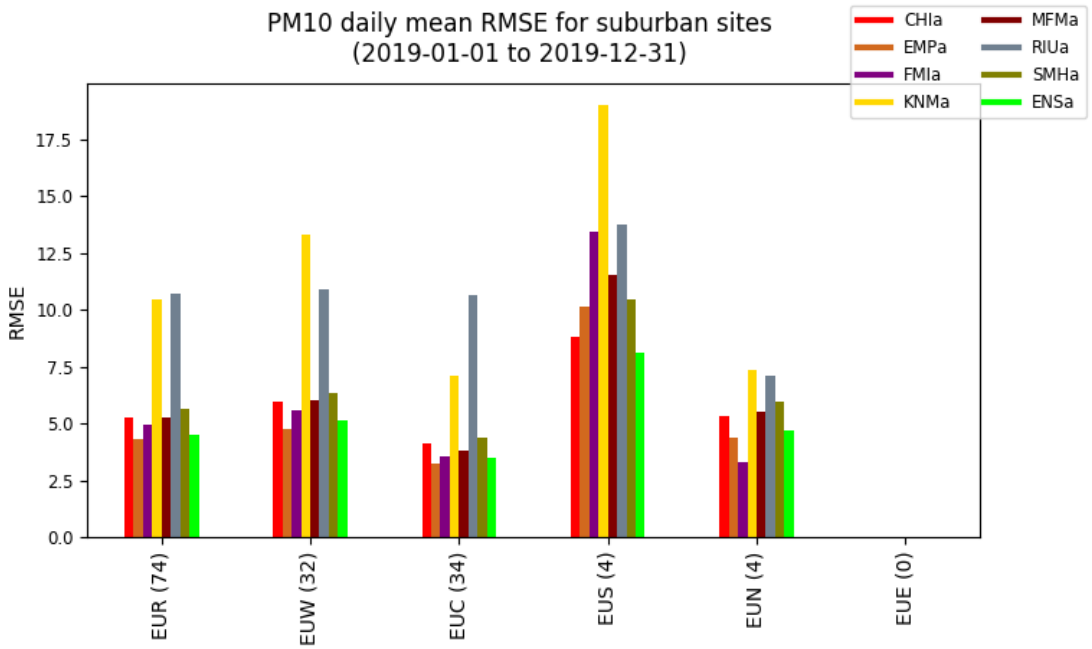


(c)

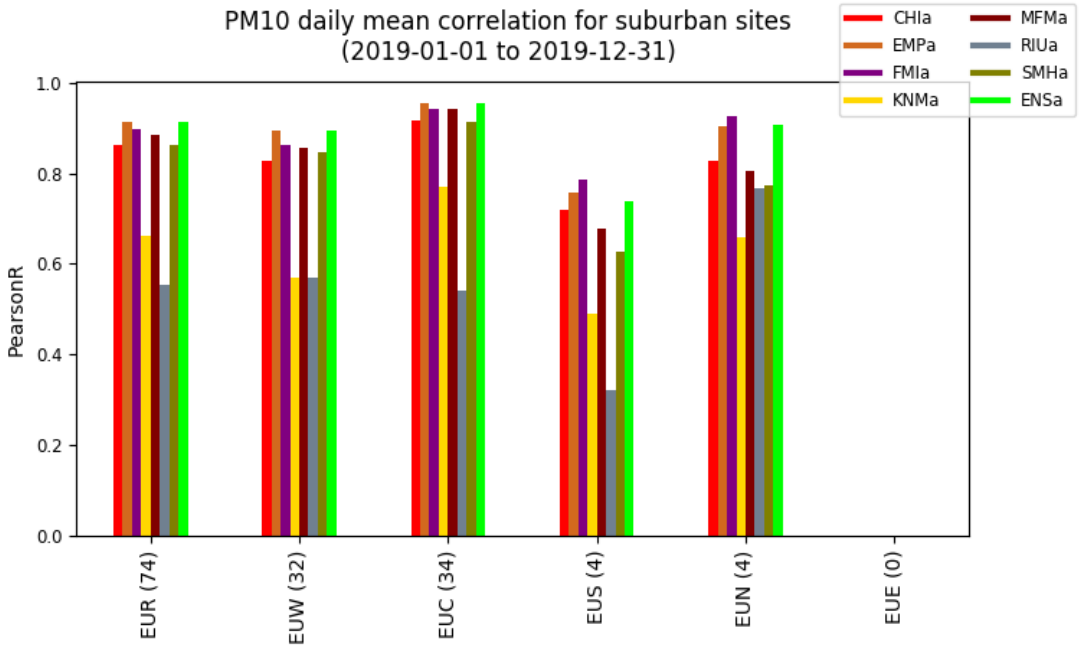
Figure 12 - CAMS Regional interim reanalyses for predicting PM_{10} daily average over the year 2019 throughout European sub-regions: Bias (a) RMSE (b), Correlation coefficient (c) at rural stations.



(a)

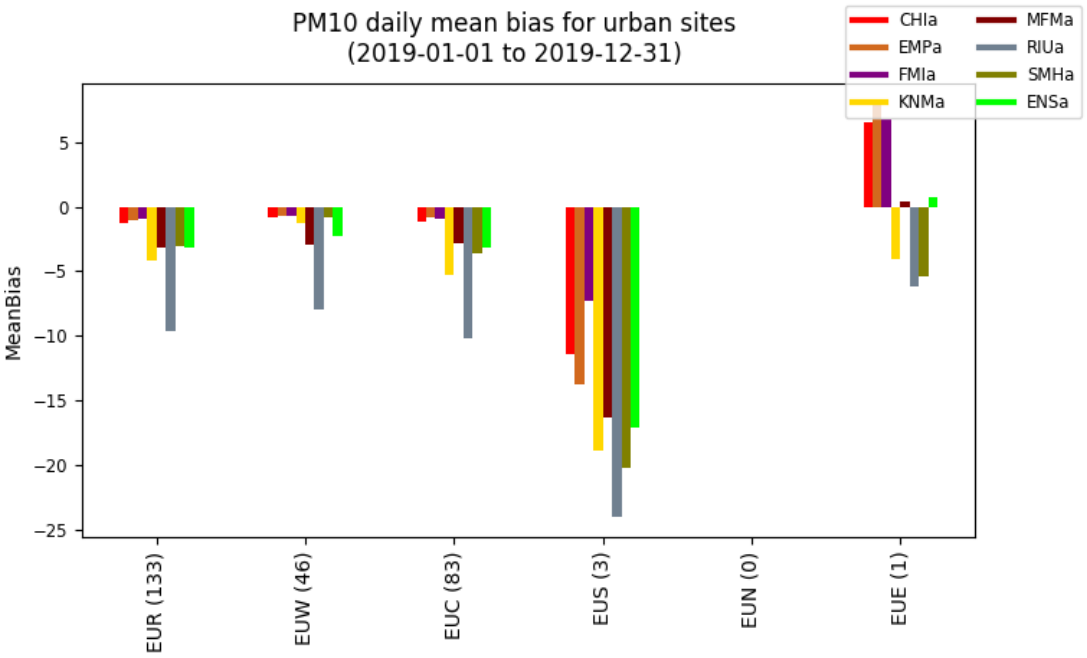


(b)

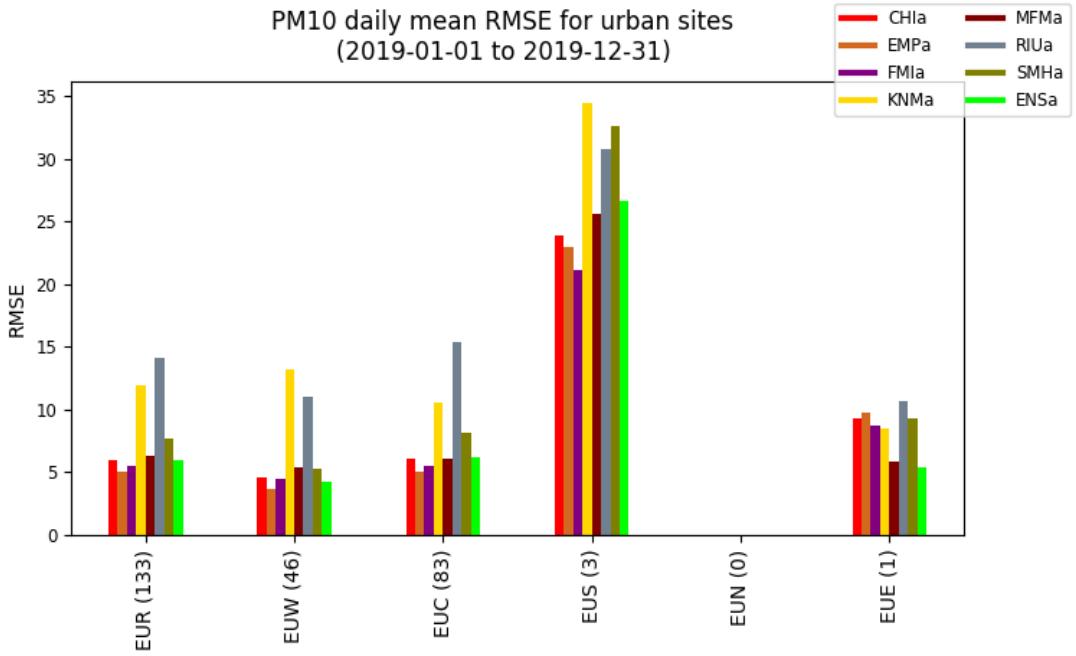


(c)

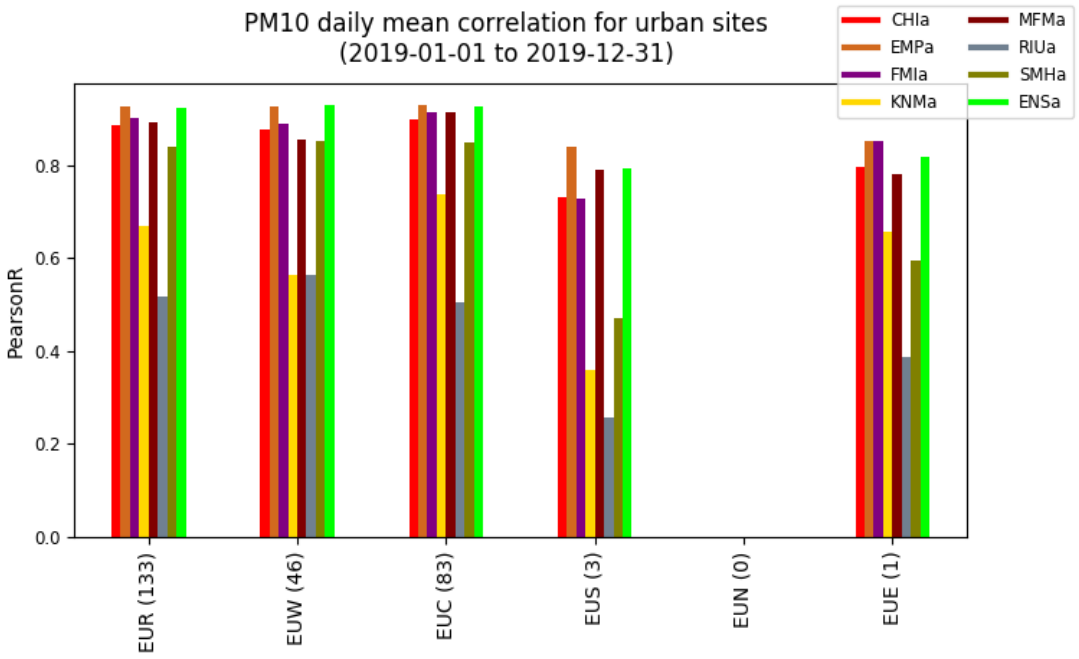
Figure 13 - CAMS Regional interim reanalyses for predicting PM_{10} daily average over the year 2019 throughout European sub-regions: Bias (a) RMSE (b), Correlation coefficient (c) at suburban stations.


PM10 daily mean bias for urban sites
(2019-01-01 to 2019-12-31)
CHIa
EMPa
FMIa
KNMa
MFMa
RIUa
SMHa
ENSa
MeanBias
5
0
−5
−10
−15
−20
−25
EUR (133)
EUW (46)
EUC (83)
EUS (3)
EUN (0)
EUE (1)

(a)

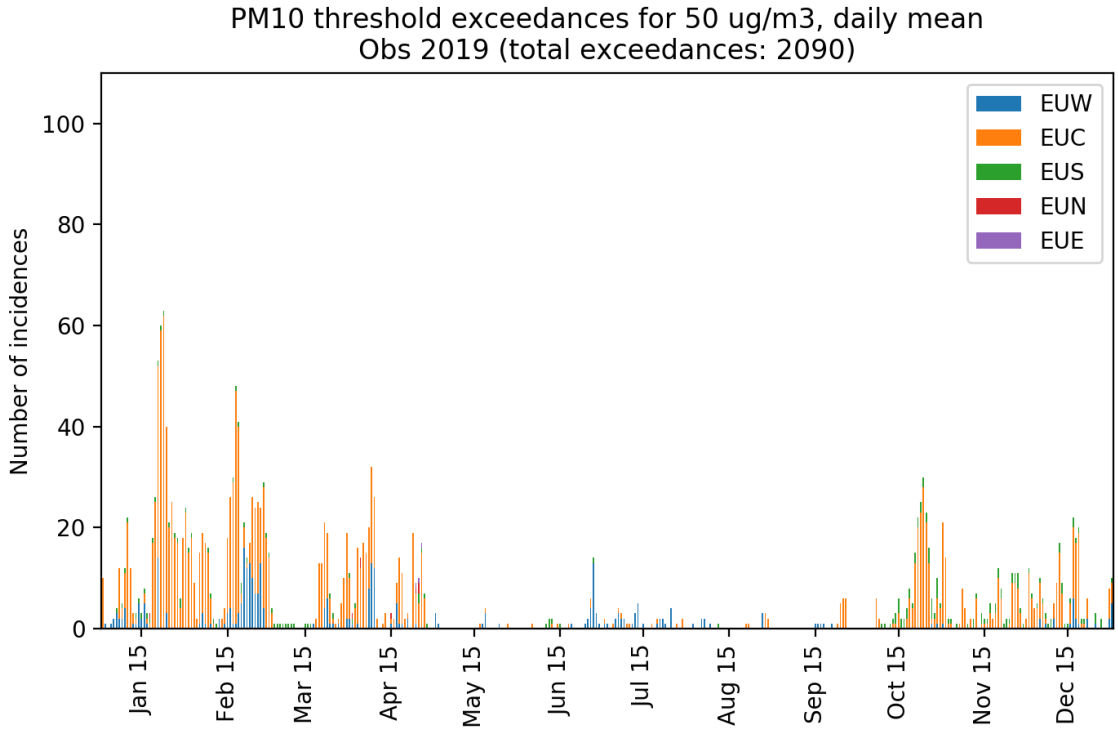

PM10 daily mean RMSE for urban sites
(2019-01-01 to 2019-12-31)
CHIa
EMPa
FMIa
KNMa
MFMa
RIUa
SMHa
ENSa
RMSE
35
30
25
20
15
10
5
0
EUR (133)
EUW (46)
EUC (83)
EUS (3)
EUN (0)
EUE (1)

(b)



(c)

Figure 14 - CAMS Regional interim reanalyses for predicting PM_{10} daily average over the year 2019 throughout European sub-regions: Bias (a) RMSE (b), Correlation coefficient (c) at urban stations.



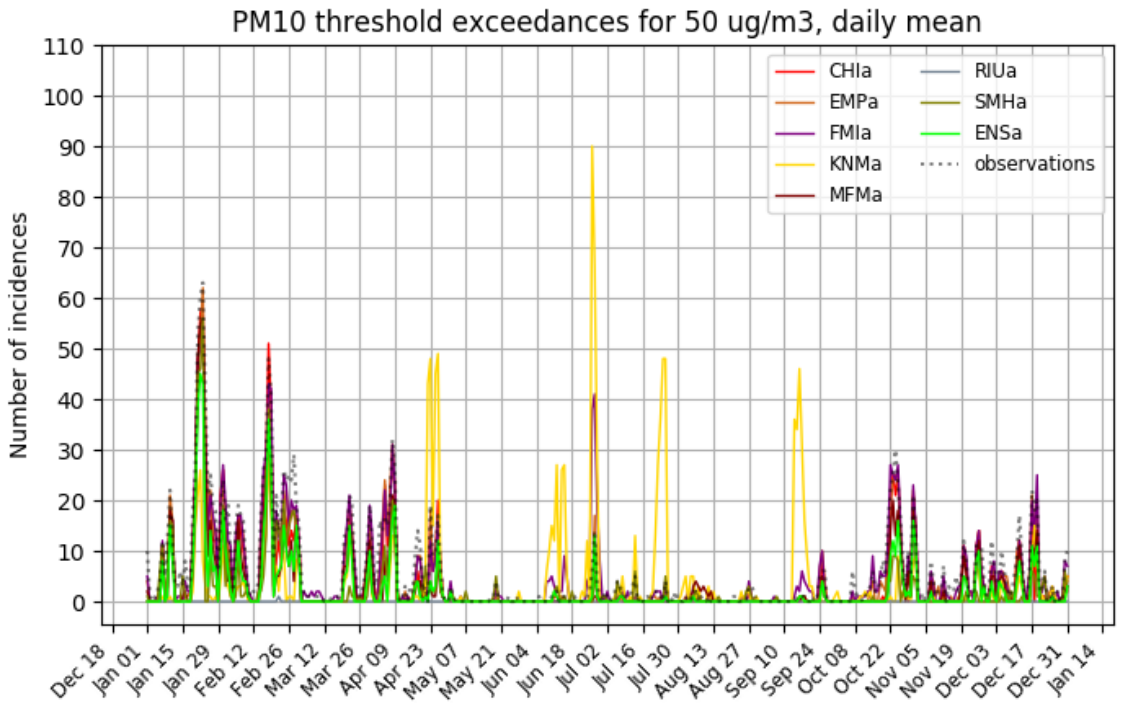


Figure 15 - Number of exceedances of daily limit value for PM_{10} in 2019 – observed (top) and modelled by interim reanalyses (bottom).

Figure 15 shows the number of exceedances of the PM_{10} daily limit value (50 µg/m^3), sorted per region and which mainly occurred during the first quarter of the year. Both observed and re-analysed data are presented and compared.

Most of the reanalyses are able to capture PM_{10} pollution episodes at the right time and with the correct duration, even if some false alarms occurred like for KNMa during the summer months.

ENSEMBLE shows interesting skill on the timeseries of the threshold exceedances. However, when focusing on Figure 16, we can note that only ~40 % of the exceedances are well captured whereas the best individual reanalyses manage to capture more than 60 % of the exceedances. However, it is worth noting that the ENSEMBLE makes a low number of false alarms referring to its high success ratio, meaning that ENSEMBLE reanalyses indicate exceedances with a very good level of confidence.

Other reanalyses show a large panel of responses in terms of ability to detect threshold exceedances, all with a tendency to underestimate threshold exceedances (FBIAS < 1 due to more missed events than false alarms). Some have very poor scores due to their chronic underestimation of the PM_{10} concentrations. Other manage to capture well part of the exceedances but with also many false alarms; this is problematic as it leads to describing as polluted, areas which are not.

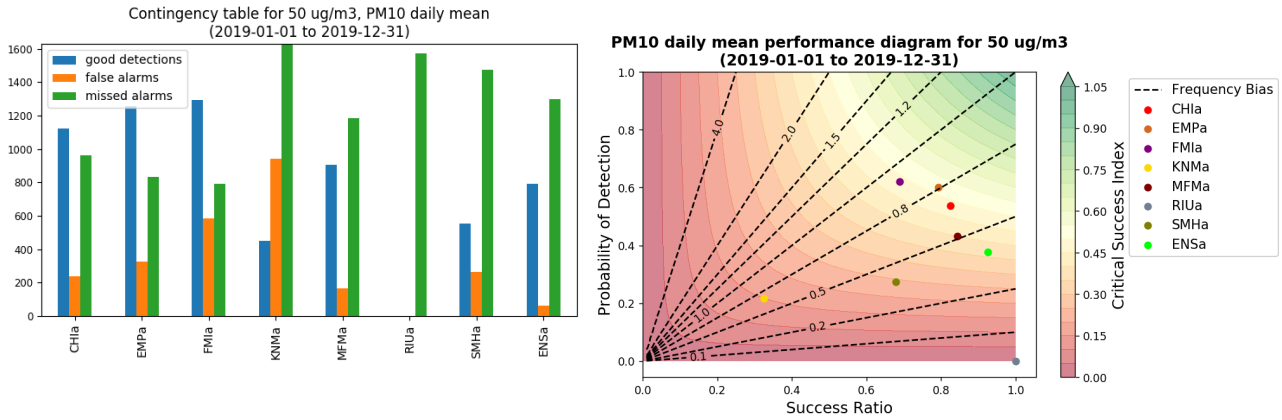


Figure 16 - Number of exceedances of daily limit value for PM_{10} in 2019 modelled by interim reanalyses.

5. Performance indicators for $PM_{2.5}$

Figure 17 shows the Taylor diagram obtained for $PM_{2.5}$ daily averages over the year 2019, for CAMS Regional individual and ENSEMBLE reanalyses. Two distinct groups appear on the plot with performances that are not too far, with correlation between 0.8 and 0.9 and RMSE between 4 and 5 µg/m^3. One model is far away from these values, with much lower performances.

The evaluation of models' performances for $PM_{2.5}$ was constrained by the low number of stations available. This limit is clearly highlighted considering the maps on Figure 18. However, where some measurements are available, the results are rather good: bias is around 0 for most of the European countries except for some stations in Poland, Serbia, Greece and Italy that also have the highest RMSE and the poorest correlations. Correlation coefficient exceeds 0.8 everywhere else. RMSE stays generally below 5 µg/m^3. Even if some concerns about the representativeness of these scores can be raised considering the low number of stations, we can consider those figures as encouraging. The values are remarkably homogeneous regarding the geographical location of the stations.

Those conclusions are confirmed by the analyses of the histograms by sub regions, showing correlation coefficient and RMSE estimated for each model and for the various station typologies (rural and urban respectively on Figure 19 and Figure 20). Bias and RMSE generally have values higher than for PM_{10} and the model responses show a larger variability. The statistical scores are quite satisfactory, with correlation coefficient generally higher than 0.8 and RMSE generally lower than 10 µg/m^3. As for PM_{10}, the ENSEMBLE has the best performances for $PM_{2.5}$.

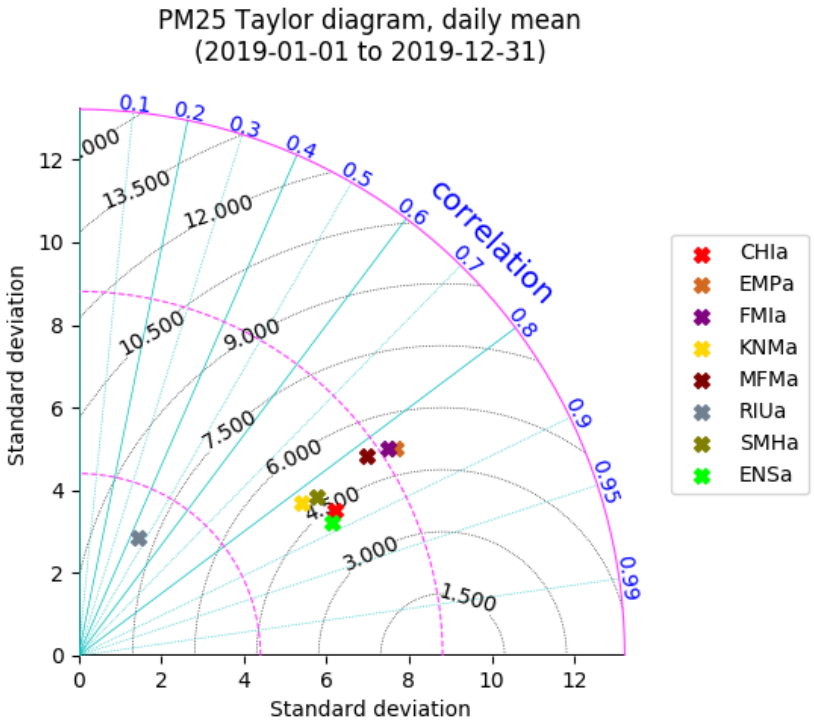
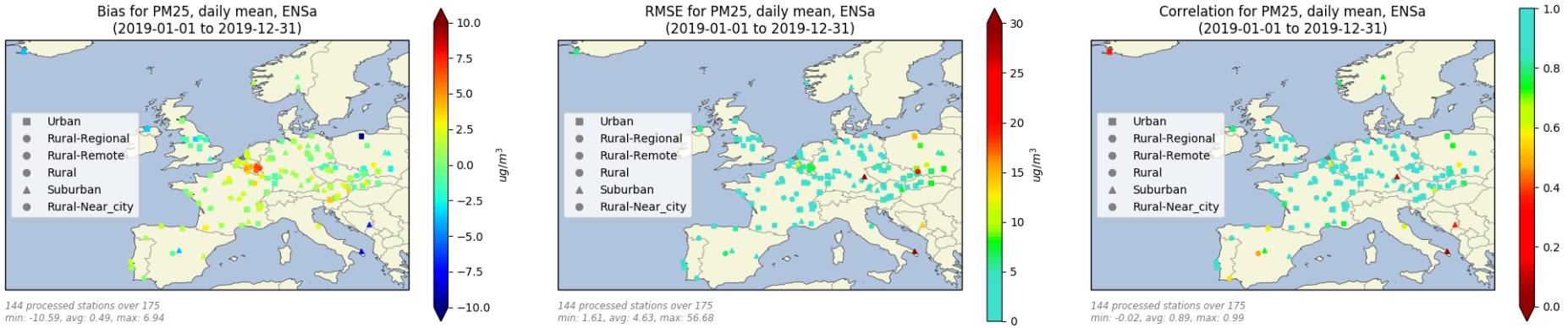
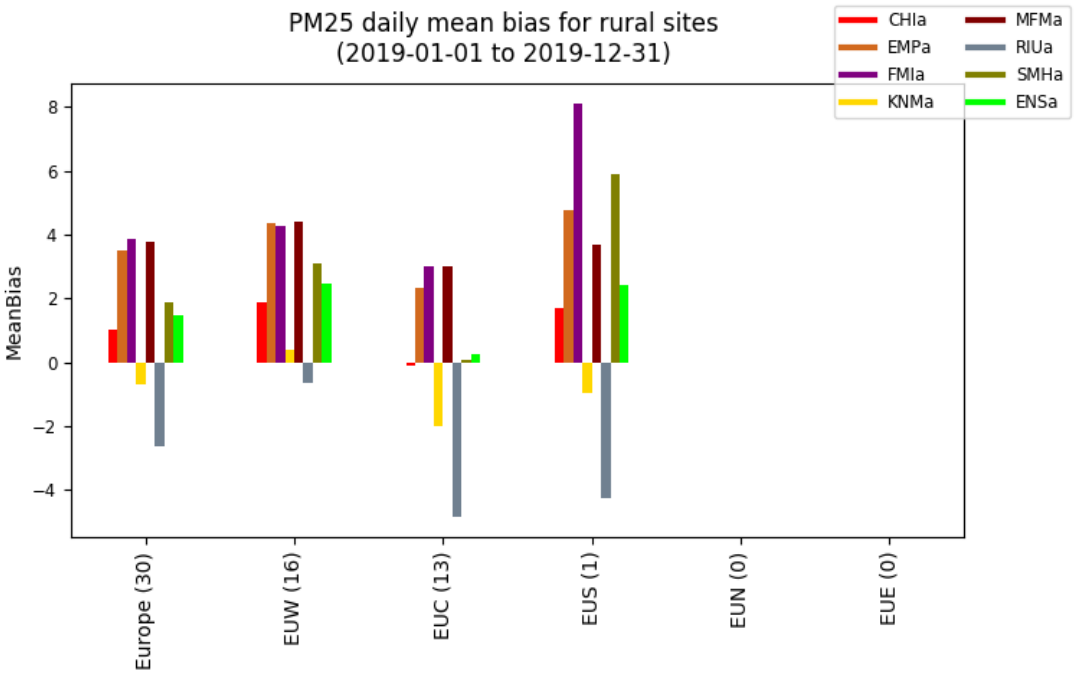


Figure 17 - Taylor diagram presenting the performances of the CAMS Regional ENSEMBLE interim reanalyses to predict PM_{25} daily average in 2019.

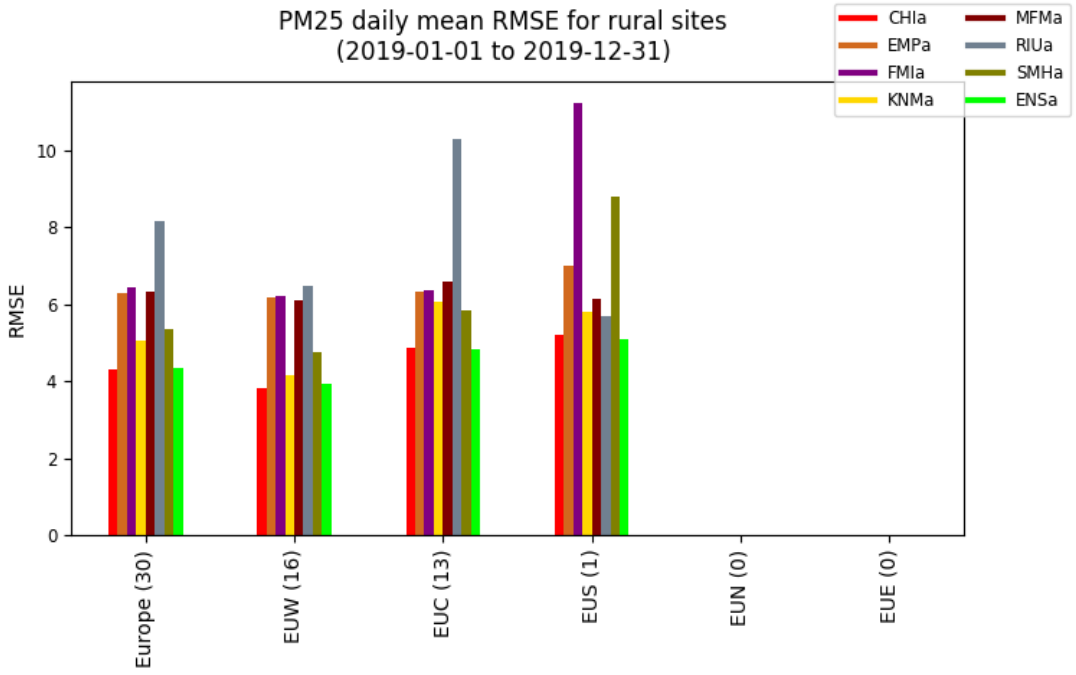


(a) (b) (c)

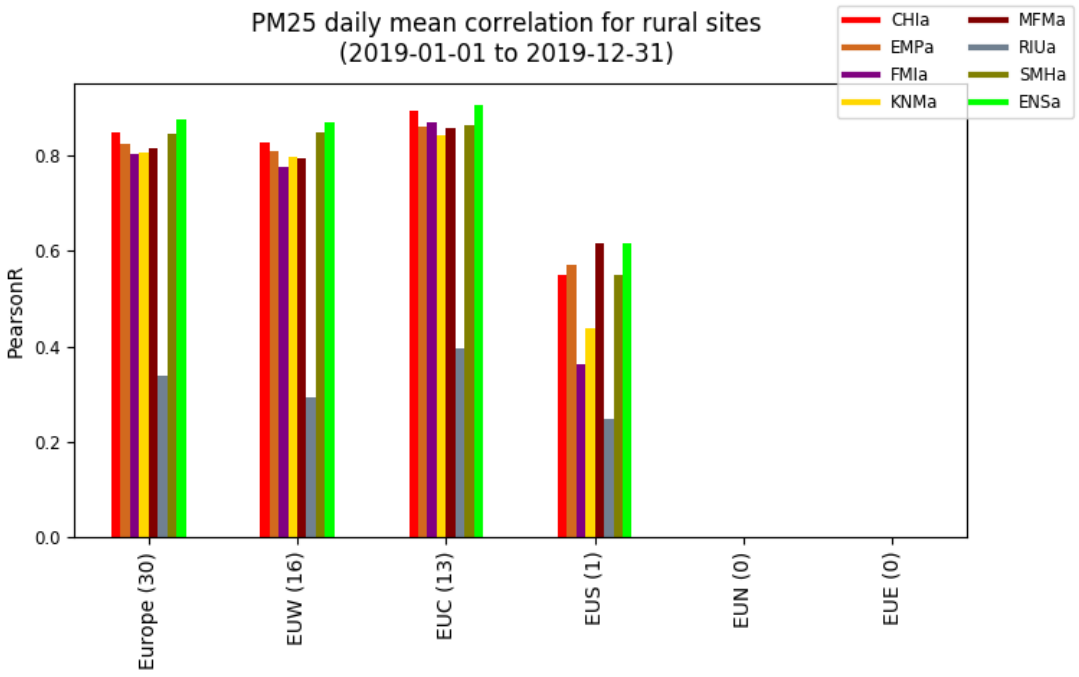
Figure 18 - Maps of Statistical scores of the ENSEMBLE interim reanalyses results against the observation validation dataset from the AQ e-reporting database for the $PM_{2.5}$ daily average over the year 2019: Bias (a) RMSE (b), Correlation coefficient (c).


PM25 daily mean bias for rural sites
(2019-01-01 to 2019-12-31)
CHIa
EMPa
FMIa
KNMa
MFMa
RIUa
SMHa
ENSa
MeanBias
8
6
4
2
0
−2
−4
Europe (30)
EUW (16)
EUC (13)
EUS (1)
EUN (0)
EUE (0)

(a)

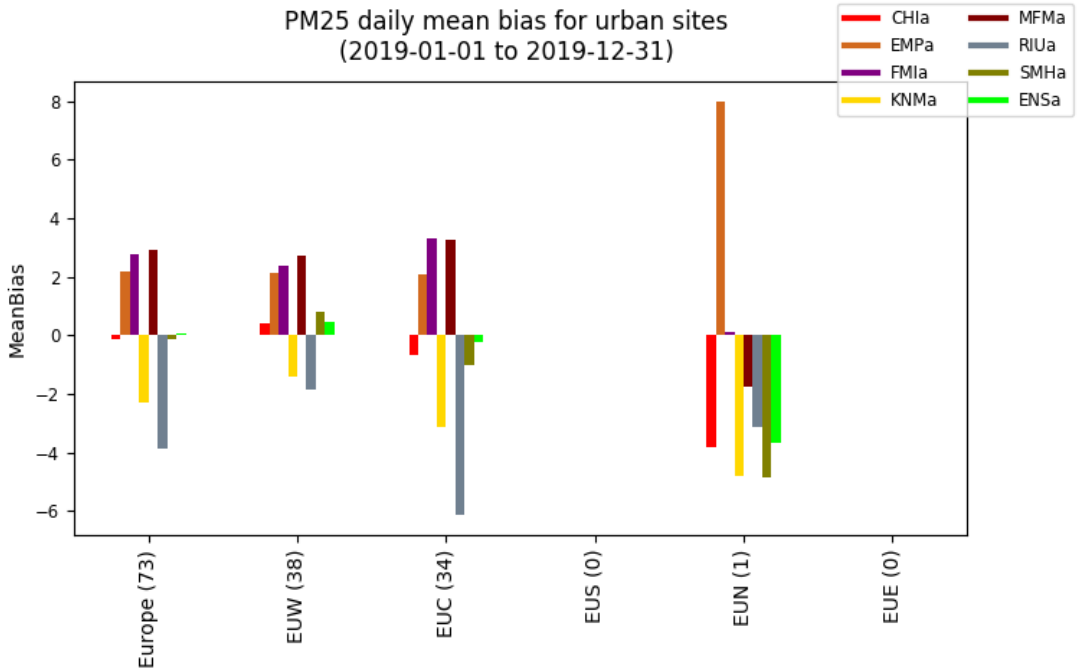

PM25 daily mean RMSE for rural sites
(2019-01-01 to 2019-12-31)
CHIa
EMPa
FMIa
KNMa
MFMa
RIUa
SMHa
ENSa
RMSE
10
8
6
4
2
0
Europe (30)
EUW (16)
EUC (13)
EUS (1)
EUN (0)
EUE (0)

(b)

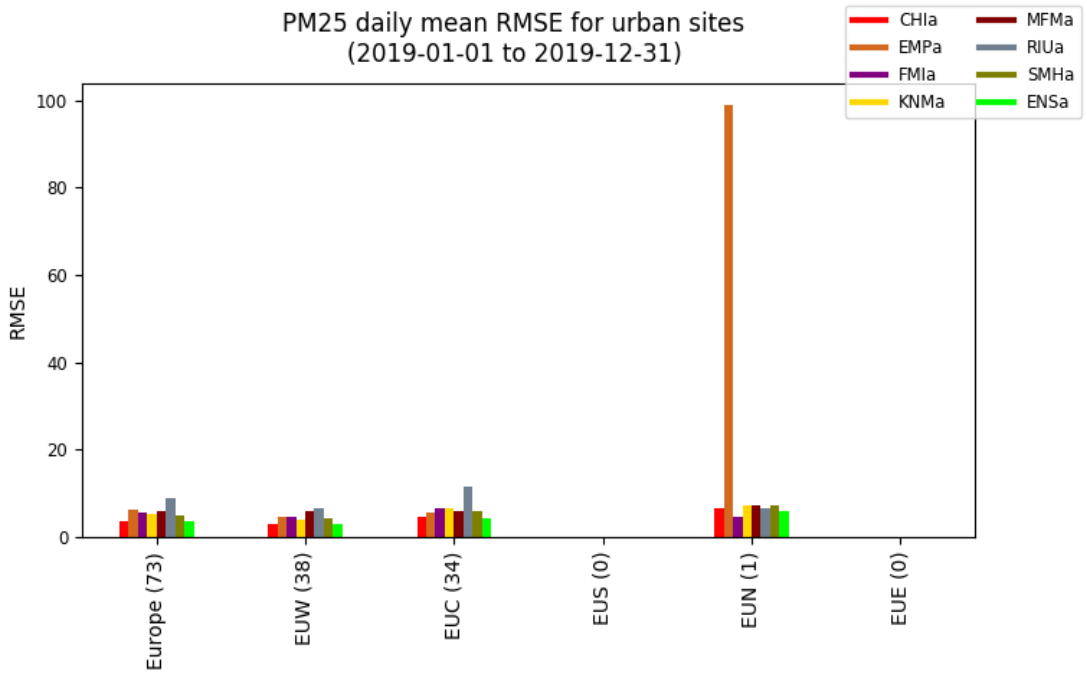


(c)

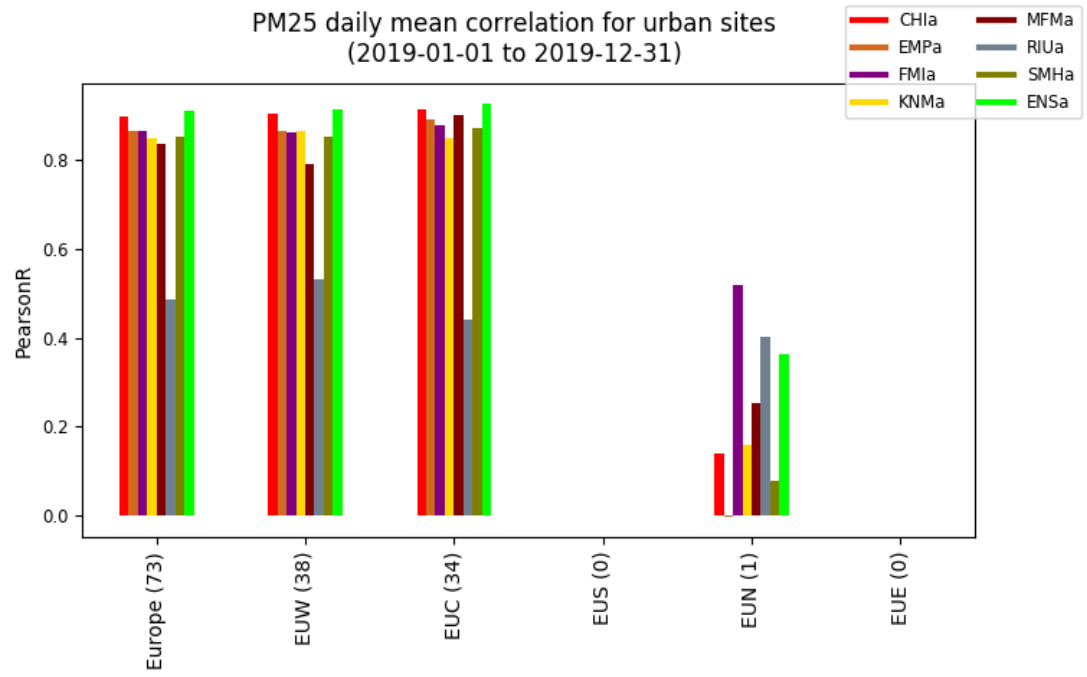
Figure 19 - CAMS Regional interim reanalyses for predicting $PM_{2.5}$ daily average over the year 2019 throughout European sub-regions: Bias (a) RMSE (b), Correlation coefficient(c) at rural stations.



(a)



(b)



(c)

Figure 20 - CAMS Regional interim reanalyses for predicting $PM_{2.5}$ daily average over the year 2019 throughout European sub-regions: Bias (a) RMSE (b), Correlation coefficient (c) at urban stations.

6. Performance indicators for SO_2

Stations measuring SO_2 throughout Europe are sparse and the background level of concentrations is well under the limit values (except when a volcano eruption occurs, which was not the case for this year). Exceedances are essentially located nearby industrial areas.
The Taylor diagram (Figure 21) shows a large variability of the model responses, with a similar and poor correlation of 0.2 for all and a RMSE ranging from 4 to 6 µg/m^3.
For most of the European countries, the ENSEMBLE has bias close to 0 (Figure 22). Some urban stations located in Belgium, in Czech Republic and along the coastline in Italy have very high RMSE. Almost all stations depict very poor correlations. The correlation is usually not very good, because of the low concentrations of SO_2 over background stations which variations are complex to simulate.

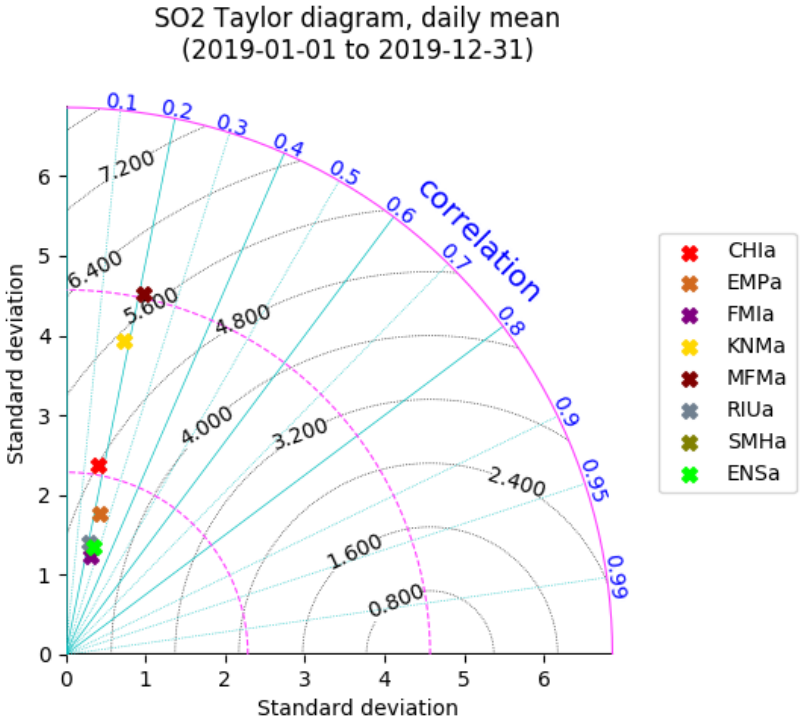
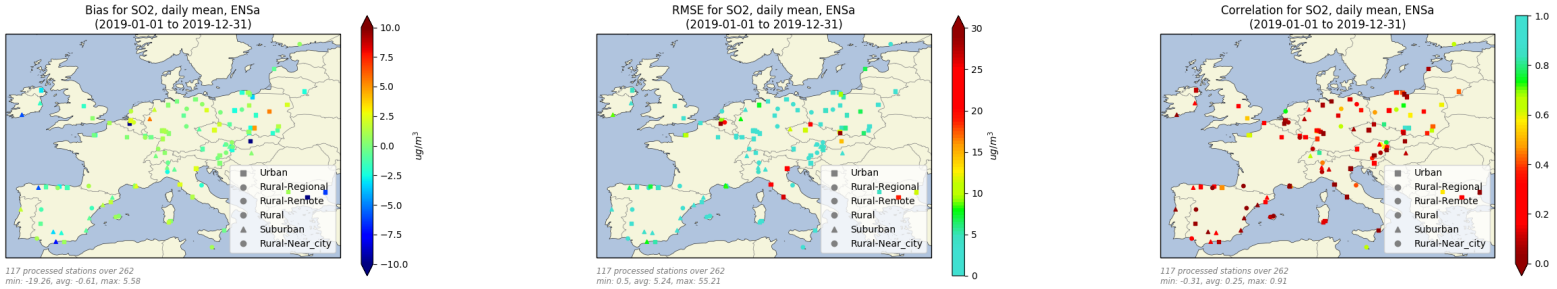


Figure 21 - Taylor diagram presenting the performances of the CAMS Regional ENSEMBLE interim reanalyses to predict SO_2 daily average in 2019.



(a) (b) (c)

Figure 22 - Maps of Statistical scores of the ENSEMBLE interim reanalyses results against the observation validation dataset from the AQ e-reporting database for the SO_2 daily average over the year 2019: Bias (a) RMSE (b), Correlation coefficient (c).

Conclusion

The present report provides an analysis of the performances of the interim air quality reanalyses throughout Europe, produced by the CAMS Regional service for the year 2019. It focuses on ENSEMBLE air quality reanalyses resulting from the combination of seven well-validated and documented chemistry-transport models results. We call here "interim" reanalyses data assimilated fields of air pollutant concentrations based on up-to-date observation data. Because such data is quickly available after their production, the validation process it is submitted to is not necessarily achieved and the data should be considered as "interim" data. Nevertheless, we found interesting to elaborate interim reanalyses as first guess of air pollution patterns and levels that developed in Europe in 2019. Such information can be used to support Member States for the regulatory reporting duty on air quality (according to Directive 2008/50/EC). This is the reason why it is important to carefully evaluate the simulations against observations that are not used for the reanalyses production.

INERIS run this process and computed a number of performance indicators and scores for ozone, nitrogen dioxide, sulfur dioxide, PM_{10} and $PM_{2.5}$ concentrations. They are presented in this report using maps, Taylor diagrams and histograms. The main conclusions arising from this analysis are the following:

- As for the previous years, too little up-to-date observation data was available to perform an extensive evaluation of interim reanalyses over the whole of Europe (except for Central and Western Europe). Very few observations can be available in Eastern Europe and also in Southern and Northern European regions that are not correctly covered. This is frustrating since they correspond to areas where there are more uncertainties (especially because of emissions).
- In Western and Central Europe, where there are more stations for the evaluation of the models' performances, results are generally more representative and correct. The quality of the reanalyses is generally similar to the previous years, as slight improvement was noticed compared to 2018 about the O_3 and PM_{10} scores.
- The European Environment Agency is building capacity to strengthen quality assurance procedures in the coming years and more countries are supposed to deliver up-to-date data, which will impact positively the interim reanalyses production process.
- For all pollutants, the performances are always of lower quality than what can be achieved with the validated reanalyses process, for which more stations are available and observation datasets are validated.
- The ENSEMBLE reanalyses give the best results for ozone when focusing on classical statistical scores. Ozone daily maxima are generally underestimated. Correlation coefficient ranges between 0.8 and 0.9 and RMSE around 10 $\mu g/m^3$ at rural and suburban locations. However, when looking at the threshold exceedances, it is worth noting the low capabilities of the ENSEMBLE reanalyses to detect concentrations above the standards (25%) and its good skills to keep the number of false alarms at a very low level. This highlights the high confidence associated with the ENSEMBLE's exceedances represented on maps.

- Excepting one model (RIUa), the model responses for ozone are very close and slightly better than in 2018 for bias, RMSE and correlation. A bigger diversity of responses appears when considering the capability of detection of the threshold exceedances.
- The performances of nitrogen dioxide ENSEMBLE reanalyses are quite stable with previous years and with satisfactory scores, but lagging behind the performances of the other pollutants (without considering SO_2). RMSE is around 12 µg/m^3, bias shows an underestimation of 10 µg/m^3 and correlation is close to 0.7. One model behaves as outliers (SMHa).
- PM_{10} is the pollutant for which model responses range in a large interval. Two models (RIUa and KNMa) are aside of a group (including the ENSEMBLE) where correlation coefficient ranges from 0.8 to 0.9 and RMSE from 4 to 6 µg/m^3, depending on the model and the station typology. Model responses have a bias close to 0 with a tendency to become slightly negative for urban stations. Anyway, the ENSEMBLE shows good performances, better than for the previous year. Part of this result might be explained by the scores over Polish stations which improved this year. The homogeneity of the ENSEMBLE's scores whatever the typology of stations considered is also noticed.
- Moreover, the evaluation demonstrates how the Ensemble approach, based on a median average of involved models is not appropriate to simulate exceedances of threshold values. Only 35% of good detection of exceedances of the PM_{10} daily limit values was correctly caught by the ENSEMBLE, whereas the best reanalyses got 60 %. As for ozone, the ENSEMBLE reanalyses produce a very low number of false alarms.
- Despite only few $PM_{2.5}$ measurement data was available for the evaluation, the results obtained for this pollutant are promising. The individual models' responses are quite consistent, and the Ensemble median gives the best results. Correlation coefficient is close to 0.9 and the RMSE between 4 and 5 µg/m^3, which is good. Once again, the conclusions are limited by the low number of stations available in some geographical areas and should be consolidated and improved in future interim assessments, when the up-to-date data gathering process at the EEA is strengthened.
- The model representativeness is limited to correctly reproduce SO_2 concentrations in Europe, due to the characteristics of the emissions sources of such pollutant. This is illustrated by the results that show low RMSE and bias due the low background concentrations measured and very poor correlation for almost all the European stations, highlighting the complexity for the model to reproduce the temporal variability of the concentrations.

Copernicus Atmosphere Monitoring Service

ECMWF - Shinfield Park, Reading RG2 9AX, UK

Contact: info@copernicus-atmosphere.eu

atmosphere.copernicus.eu copernicus.eu ecmwf.int